Mapping quantitative trait loci in microbial populations

Sayanthan Logeswaran

PhD

The University of Edinburgh 2010

Abstract

Linkage between markers and genes that affect a phenotype of interest may be determined by examining differences in marker allele frequency in the extreme progeny of a cross between two inbred lines. This strategy is usually employed when pooling is used to reduce genotyping costs. When the cross progeny are asexual the extreme progeny may be selected by multiple generations of asexual reproduction and selection. In this thesis I will analyse this method of measuring phenotype in asexual cross progeny. The aim is to examine the behaviour of marker allele frequency due to selection over many generations, and also to identify statistically significant changes in frequency in the selected population. I will show that stochasticity in marker frequency in the selected population arises due the finite initial population size. For Mendelian traits, the initial population size should be at least in the low to mid hundreds to avoid spurious changes in marker frequency in the selected population. For quantitative traits the length of time selection is applied for, as well as the initial population size, will affect the stochasticity in marker frequency. The longer selection is applied for, the more chance of spurious changes in marker frequency. Also for quantitative traits, I will show that the presence of epistasis can hinder changes in marker frequency at selected loci, and consequently make identification of selected loci more difficult. I also show that it is possible to detect epistasis from the marker frequency by identifying reversals in the direction of marker frequency change. Finally, I develop a maximum likelihood based statistical model that aims to identify significant changes in marker frequency in the selected population. I will show that the power of this statistical model is high for detecting large changes in marker frequency, but very low for detecting small changes in frequency.

Declaration

I declare that this thesis has been composed by myself and is entirely my own work, except for the input mentioned below. The work contained within this thesis has not been submitted for any other degree or professional qualification.

Chapter 4: The malaria data that was used in the analysis in Section 4.3.2 was generated by members of Professor Richard Carter's group.

Acknowledgments

I would like to thank my supervisor Nick Barton for being extremely helpful, supportive and accessible throughout the previous four years.

I would also like to thank members of Richard Carters Lab, particular Naomi Gadsby and Sandie Cheesman, for many helpful discussions. I also thank Ian White for statistical advice.

Finally, I am extremely grateful to the Marie Curie Early Stage Research Training Programme for funding this research.

Contents

Chapter 1: General Introduction

1.1 Phenotype Variation

Phenotypic variation in a population comes in many forms and may be influenced by many different factors. In the simplest case the variation in a trait may be due to just a single genetic locus. These traits are usually referred to as simple or Mendelian traits, where the alleles present at that single locus gives rise to most of the phenotypic variation. Traits that are influenced by more than one locus are usually referred to as quantitative traits. These are perhaps the most common type of traits and the genetic architecture of these traits is generally much more complicated. Despite this complexity the benefits in deciphering the genetic basis of the variation in quantitative traits is manifold. In human populations, the most obvious rewards come from understanding complex diseases such as heart disease, diabetes, and so on. An understanding of the genetic basis of these diseases may help towards development of drugs to overcome these disorders or predict disease susceptibility. In others areas such as agriculture, production can be improved by understanding the genetic basis of variation in crops yield, milk production, and meat quality. So clearly these traits have health and economic importance, and so elucidating the genetic basis of these would be very beneficial.

1.2 Genetic Basis of Phenotypic Variation

For any particular trait, we firstly need to know how many loci give rise to most of the phenotypic variation and what effect each locus has on the trait. Some traits appear to be influenced by a very large number of loci, each having a very small effect on the phenotype. This appears to be the case in many human traits. An example would be

height where studies have identified at least 40 loci that affect the trait, but the combined effect of these loci only account for about 5% of the variance (Manolio *et al.*, 2009). Other examples include human disease traits such as Crohn's disease, where 32 loci have been identified which account for 20% of the genetic variance (Barrett *et al.*, 2008). In other studies, however, relatively fewer loci explain most of the genetic variation in trait value. Typically, a single locus will have a very large effect on phenotype and contribute towards most of the genetic variance in trait value, and a number of other loci would be involved with each contributing a small amount to the genetic variance. An example of this can be seen with studies of chemical resistance traits in yeast (Ehrenreich *et al.*, 2010). Here, for one particular chemical resistance trait, 9 loci were identified which accounted for 70% of the genetic variance, where a single locus accounted for 40% of the variance and the other 8 loci each accounted for less then 10% of the genetic variance. So, the distribution of the number and effect of loci affecting a quantitative trait is not obvious, and can differ widely between traits.

Interactions within and between loci must be characterized. Interaction between alleles within a single genetic locus is known as dominance, and interaction between loci is known as epistasis. The consequences of these interactions are that the effect that an allele has on phenotype is dependent on the other allele(s) present at that particular locus (dominance) and/or alleles present on other loci (epistasis). Examples of epistatic interactions have been observed in many studies, with examples from yeast (Sinha *et al.*, 2006), plants (Kroymann & Mitchell-Olds, 2005), mice (Brockmann *et al.*, 2000) , and flies (Yamamoto *et al.*, 2009). These interactions generally make identification of loci more difficult. For example, studies have shown that the effect that a locus has on phenotype may be completely masked by the presence of alleles at other loci. An example of this can be seen with coat color in pigs (Carlborg & Haley, 2004). Alleles at a certain locus confer a dark color on pigs. However, if a particular dominant allele is present at another locus, then the pigs are white. So, the darkening effect of the alleles at one locus is completely masked by the presence of the alleles at a different locus. Other studies have shown that interaction can be difficult to detect due to the large number of

pairwise (or *n*-wise) possibilities. For example, studies have shown that loci could have individual negligible effects on phenotype, but may have significant effects when combined with other loci of negligible effects (Carlborg *et al.*, 2003).

Once a locus has been established to have some effect on phenotype one must determine which DNA variant(s) causes the change in phenotypic value. As there may be several DNA variants that uniquely determine a particular allele, one must determine which of these variants has an effect and what that variants effect is. For instance, in a study aimed at mapping loci responsible for sporulation efficiency in yeast, three biallelic loci were identified to have an effect on the trait (Deutschbauer & Davis, 2005). There were between 5 and 20 single nucleotide polymorphisms (SNPs) that distinguished the alleles at these three loci. However, only one SNP in each of the three loci was responsible for the differences in trait value. This contrasts with other studies where many variants within a single locus have found to have an affect on phenotype. An example is a locus found to have an association with human lupus disease (Graham *et al.*, 2007). Here haplotypes of three variants within a single locus were shown to have at least three distinct levels of risk to the disease. The three variants included 2 SNPs and a 30bp insertion/deletion. So, it can be seen that the numbers and type of variant(s) within a locus that affect the phenotype can be very different. Also, these polymorphisms need not be non-synonymous changes in protein coding regions, but can be synonymous changes, and also these variants often occur in non coding regions such as introns (Flint & Mackay, 2009).

Establishing a full understanding of each of the above items is a difficult process, which is made even more difficult by the fact that for any single trait any of the above could change with environment or sex. In studies in Drosophila (Dilda & Mackay, 2002) and mice (Vaughn *et al.*, 1999) many loci have shown to have sex specific effects. That is, the effect of a locus on phenotype may change in magnitude or direction, or may not be present at all in the other sex. Similar results have been shown with environmental changes, where loci found to have an effect in one environment show no effect in other

environments. An example of this can be seen with the mutations responsible for melanism in pocket mice (Nachman *et al.*, 2003). In this study four mutations at a single locus were found to be responsible for melanism in a particular population. However, in another population of pocket mice, the four mutations showed no association with melanism, indicating that different loci were responsible for melanism in different populations.

So, it can be seen that deciphering the complete pathway from genotype to phenotype for most quantitative traits is quite a tall order. In most studies, the initial step is to scan the genome and search for general chromosome regions that may affect the phenotype. Any regions identified by the genome scan will then be examined further to identify individual loci that may affect the trait. The two general strategies that are usually used to do this initial genome scan are association studies and linkage analysis. Both methods essentially seek to capture correlations between known marker loci and the phenotype of interest. Linkage analysis attempts to achieve this by using controlled crosses or small families, whereas association studies rely on linkage disequilibrium in a population of unknown pedigree. This thesis concentrates on linkage analysis and its application to a pooling technique in microbial populations. Outlined next will be the basics of linkage analysis and its limitations, followed by a brief outline of pooling in linkage analysis experiments and its recent application to microbial populations.

1.3 Linkage Analysis

The aim of linkage analysis is to identify regions of the genome that influence a particular trait. These regions are typically referred to as quantitative trait loci (QTL). QTL don't necessarily refer to any particular genetic locus, but rather a general stretch of chromosome that may contain a gene (or several genes) that influence a particular trait. The general strategy to identify QTL in linkage analysis experiments is to look for associations between known marker alleles and the phenotype (Sax, 1923; Thoday,

1961). In most experiments this is achieved through a series of genetic crosses. Observing how the marker alleles segregate in a cross and correlating these observations with the corresponding phenotype of the cross progeny can be used to detect QTL. There are many variations to this type of methodology (Darvasi, 1998; Lynch & Walsh, 1998) so outlined below are the basic steps.

In general linkage analysis experiments are initiated with a cross between two inbred lines. These two parental lines will differ in value for the trait being analysed, typically one line representing a high trait value and the other a low trait value. Both lines would usually be homozygous at all loci and are genetically distinguished by a set of polymorphic markers. The aim is to generate a series of crosses from these parental lines so that the progeny would have a distribution of trait values and whose genotypes are a unique mixture of the parental markers. This can be achieved after two crosses (one if organisms are haploid). In the cross progeny, associations between markers and the trait can be analysed to identify QTL. In order to do so each cross progeny is genotyped for the marker alleles they contain and scored for their phenotypic value. Individuals carrying markers that are physically linked to alleles that affect the trait should show statistical correlations with the trait value. Hence, any marker loci that show statistically significant correlations with trait value can be assumed to be linked to one or more alleles that affect the trait. Since the position of the marker alleles on the genome are known, it is possible to define a general region for the QTL. There are a wide variety of statistical methods that exist that try and identify these marker phenotype correlations and define QTL regions (Broman, 2001).

1.4 Limitations of Linkage Analysis

In terms of fully deciphering the genetic basis of complex traits, the above procedure is just an initial step and the predictions that come from it are, in most cases, far from conclusive. The two main issues that arise are, underestimating the number of loci

involved, and difficulties in identifying the actual genetic locus (or loci) that affect the trait within a predicted QTL.

The first of these limitations happen when many loci influence a trait. Most experiments can only usually detect QTL of relatively large effect. This is mainly due to the size of the mapping population, as most linkage analysis experiments have sample populations that are too small to detect most small effect QTL. For example, if a marker was linked to a QTL of small effect, then that particular marker would only be associated with a small change in phenotypic value. This small deviation from the phenotypic mean would in most cases be too difficult to detect, unless a very large number of cross progeny have that particular marker and show the same effect. With large effect QTL much fewer progeny are needed as the large deviation from the phenotypic mean can be easily recognized. So, unless very large sample sizes are available, many small effect QTL would go undetected.

Another reason why standard linkage analysis experiments may miss QTL is due to tight linkage between QTL with opposing effects. For instance, if two closely linked QTL have effects in opposite directions, then the net effect of the two QTL may be very low. Due to the overall net low effect, both QTL may go undetected. This type of behavior has been frequently reported in QTL studies. An example is provided from a study attempting to identify genes responsible for high temperature growth in yeast (Steinmetz *et al.*, 2002). Here, a single QTL of large effect was identified by standard linkage analysis. On fine mapping, this apparent large effect QTL separated out into three smaller effect tightly linked QTL, where the effect of one the QTL was in the opposite direction to the other two QTL. Similar results have been reported for sporulation efficiency in yeast (Ben-Ari *et al.*, 2006) and growth rate in Arabidopsis (Kroymann & Mitchell-Olds, 2005). In most of these studies these clustered QTL were identified as their net effect was big enough to be detected as an apparent single large effect QTL, and consequently the individual smaller effect QTL were later uncovered. However, if this clustering of QTL with opposing effects is frequent in quantitative traits, then it is

conceivable there may be many QTL that go undetected as the net effect of many of these clustered regions may simply be too small to be detected.

The most problematic limitation of standard linkage analysis experiments, however, is perhaps the length of the QTL that are predicted. Ideally, the chromosome interval defining the QTL should be as small as possible so that the causative locus can be easily identified. Most standard linkage analysis experiments, however, are unable to do this and the QTL regions predicted are quite large, making it extremely difficult to pinpoint any causative genetic locus. This problem is due to the lack of recombination events in the experiment. If a block of genome surrounding an allele that affects the trait is never broken up by recombination, then markers within this block will show strong association with the trait. So, if there are only a few recombination events during the experiment then this block will be very large and potentially many markers will show this association. This results in very large QTL regions, which can contain a prohibitively large number of genes to examine. So, ideally what we would like to have is these large QTL broken up into smaller regions where each QTL is associated with only very closely linked markers. Achieving this will require much larger population sizes so that a lot more recombination events are observed.

1.5 Selective Genotyping

So, it can be seen that most standard linkage analysis experiments will only give a very crude estimation on the number and location of the QTL involved. In theory, the simplest way to address these problems is to analyse extremely large mapping populations, ensuring that more recombination events are observed. This would ensure that more small effect QTL are detected, ensure a greater probability that closely linked QTL are broken up, and enable higher precision mapping. In practice, however, this would involve analyzing population sizes at least in the thousands, and the time and costs involved in genotyping and phenotyping such large populations are usually

prohibitive. One way to offset some of these expenses would be to genotype only a smaller subset of the mapping population. This procedure is usually referred to as selective genotyping and involves only genotyping and analysing individuals based on their phenotypic value (Darvasi & Soller, 1992; Lander & Botstein, 1989). This is because in any mapping population most of the linkage information comes from individuals with extreme phenotype (Lander & Botstein, 1989). Therefore just genotyping and analyzing marker trait associations in the tails of the phenotypic distribution can reduce time and costs of the experiment. This strategy has been extensively used to map QTL in linkage analysis experiments. Some examples include QTL mapped for carcass traits in chickens (Nones *et al.*, 2006), grain and malt quality traits in barley (Ayoub & Mather, 2002), bovine ovulation rate (Kirkpatrick *et al.*, 2000), and rot resistance in sunflower (Micic *et al.*, 2005). In these studies the amount of individuals genotyped ranged from 9% to 50% of the total population, with many of the studies detecting multiple QTL of moderate to large effect. So, although these studies have shown that selective genotyping can be a cost effective and a less time consuming procedure, they still mostly have only been used to detect relatively large effect QTL and fail to resolve the general limitations of linkage analysis experiments. This again is mostly due to the sample sizes that are used. To overcome the limitations, very large numbers of extreme progeny would still need to be genotyped. In many experiments obtaining or genotyping such large numbers is usually very difficult. However, if obtaining very large numbers of extreme progeny is feasible, then one way to further reduce genotyping time and costs in these large populations is to combine selective genotyping with DNA pooling and analyse marker frequencies to detect linkage.

1.6 Marker Frequencies and DNA pooling

When analyzing marker trait associations only in the extreme progeny, such as in selective genotyping, it is possible to use changes in marker allele frequency in the extreme progeny to infer linkage between a marker and QTL (Lebowitz *et al.*, 1987).

This is because by selecting groups consisting of high and low trait values, certain markers should show differences in frequency between the two groups. For example, in an F_2 population all markers should be equally represented in the cross progeny, and thus the frequency of all markers should roughly be 0.5. However, if a group of individuals with only high trait value are chosen from the F_2 population, then there should be an abundance of markers linked to alleles that cause a high trait value, and hence an increase in frequency of these markers. Similarly, there should be a shortage of markers linked to alleles that cause a low trait value, and consequently a decrease in frequency of these markers. Markers that are completely unlinked to any locus affecting the trait should remain unchanged and have an expected frequency of 0.5. Therefore markers that show a significant difference in frequency between the two selected groups (or a significant deviation from the null expectation of 0.5) can be assumed to be linked to loci that influence the trait. Examples of using marker frequency changes to detect linkage from selective genotyping experiments can be found in various studies used to map loci in tomatoes (Foolad *et al.*, 2001; Foolad *et al.*, 2003; Zhang *et al.*, 2003).

However, this method of using change in marker allele frequency to detect linkage is mostly used when DNA pooling is employed. With DNA pooling, rather than individually genotyping each progeny in the selected group, DNA is pooled from all individuals in the selected group and marker frequencies are estimated in the pooled DNA to infer linkage. As a result of this pooling it is only necessary to genotype any particular marker once in each group, and consequently this method can reduce experimental time and costs even further than selective genotyping. This strategy is often referred to as Bulk Segregant Analysis (Michelmore *et al.*, 1991) or Selective DNA Pooling (Darvasi & Soller, 1994). It has been widely used as a relatively rapid method for detecting QTL (Quarrie *et al.*, 1999; Ruyter-Spira *et al.*, 1997; Wenzl *et al.*, 2007). Like selective genotyping it has mostly been applied to relatively smaller population sizes. However, in studies where large population sizes have been used, this pooling technique has shown it can map QTL with very high precision. An example is a study conducted by (Lai *et al.*, 2007) who used this methodology to map loci affecting

lifespan in Drosophila. Using a population size of over 21,000 F_2 flies, and analyzing marker frequencies from the DNA pooled from the extreme individuals, they mapped a total of 18 QTL. Due to the large population size, more QTL were detected than in previous studies, and the QTL intervals were much narrower, with some QTL intervals only containing a single gene.

1.7 Long Term Artificial Selection

The other main occasion when marker frequencies are used to detect linkage is in artificial selection experiments, where two lines are divergently selected (Keightley & Bulfield, 1993; Lebowitz *et al.*, 1987; Nuzhdin *et al.*, 2007). This strategy is used for quantitative traits, where the aim is to have many generations of sexual reproduction and selection. Starting from an F_2 base population, the progeny from the high and low tails of the phenotypic distribution are selected and individuals in each selected group are intercrossed to produce the next generation. This continued selection for the high and low trait values aims to produce much more extreme phenotypes than would be present in the base population, which should in turn increase the power to detect QTL. This increase in power comes from the ability the map QTL with higher precision and also the increased ability to detect significant marker frequency changes. The high resolution QTL mapping comes from the increased number of generations of recombination. More generations of recombination will ensure only tightly markers show significant changes in frequency resulting in smaller QTL intervals. Detecting frequency changes should also be easier as the continued selection for the trait should gradually fix the contributing alleles in the high and low lines. This means that there should be large differences in marker frequency between the two lines at the contributing loci. This makes the detection of smaller effect QTL easier as extremely large population sizes would not be necessary.

Despite the advantages of this methodology for mapping QTL, its use has been relatively limited. This is perhaps due to the time consuming nature of the artificial selection procedure. To generate the very extreme progeny needed for the increase in power in QTL detection, the high and low lines would need to be divergently selected for many generations. In most studies this may not be possible as the generation times may simply be too long, resulting in the experiment being prohibitively time consuming. In the limited studies that have used this procedure the results have shown that this methodology can be a powerful approach. An example is a study used to map loci affecting sternopleural bristle number in Drosophila (Nuzhdin *et al.*, 1998). In that study it was demonstrated that the effects of the QTL detected using this method are much smaller than achieved by standard linkage analysis methods.

1.8 Measuring Phenotype in Asexual Cross Progeny

In all the methods discussed above that use change in marker allele frequency to detect linkage, one must measure the phenotype of the progeny in the mapping population or in each generation of an artificial selection experiment, in order to pick out the tails of the phenotypic distribution. In most studies, the cross progeny are sexual and the phenotype is measured in standard ways. However, when the cross progeny are asexual one can use selection to measure the phenotype. Artificially selecting the asexual cross progeny over many generations is equivalent to picking out the tail of the phenotypic distribution of sexual progeny in a single generation. The longer one selects the asexual progeny (and the larger the initial population), the more extreme the tail of the phenotypic distribution that is selected.

1.9 Pooling and Asexual Cross Progeny

This method of measuring phenotype in asexual progeny and pooling has recently been used in gene mapping studies in microbes. One such method is Array Assisted Bulk Segregant Analysis (Brauer *et al.*, 2006) which has been used to map traits in yeast. Here, yeast strains differing in genetic background and trait value are crossed. The resulting asexual progeny are selected for the trait over a number of generations. A group of the selected individuals are then pooled to detect linkage. Using this methodology Brauer *et al.*, (2006) successfully mapped some major effect QTL and showed that mapping accuracy is comparable to individual genotyping, but with a reduction in experimental expenses and running time as a result of pooling.

When using this strategy in asexual cross progeny, one could also measure the phenotype directly within a pool of recombinant progeny. That is, rather than individually selecting each asexual recombinant and then pooling, one could pool the cross progeny together at the start and then select for the trait directly on this pooled progeny. The selected pool is then used to detect linkage. An example of this strategy is Linkage Group Selection (Culleton *et al.*, 2005; Martinelli *et al.*, 2005) which has been used to map loci responsible for traits in malaria parasites. Here, once again malaria parasites with differing genetic background and trait value are crossed. The resulting asexual cross progeny are pooled, and these pooled progeny are selected for the trait for many generations. Linkage is then determined by estimating changes in marker allele frequency from the selected pool. This strategy in malaria parasites has been successfully used to map mostly large effect loci. This same methodology has also been used in map loci in yeast studies. An example is a study conducted by (Segre *et al.*, 2006) where adaptive mutations in yeast were successfully mapped. However, perhaps the most successful application of this technique is a study carried out by (Ehrenreich *et al.*, 2010). In this study this pooling method in asexual cross progeny was used to map loci responsible for 17 different chemical resistance traits in yeast. They were able to use extremely large population sizes (~10^7) and as a result were able to uncover far more

loci responsible for genetic variation than most typical mapping studies. For one of the chemical resistance traits, loci explaining 70% of the genetic variance was uncovered, which included numerous small effect loci. This study illustrates the potential power of this pooling methodology for mapping loci given that extremely large population sizes are available from which large numbers of extreme phenotype can be selected and analysed.

1.10 Thesis Outline

In this thesis I will develop a basic theoretical framework for the strategy of picking out the extreme individuals in pooled asexual cross progeny by selecting for the trait over many generations. In Chapter 2, I will outline the basic model. This basic model will concentrate on the simplest genetic model of selection at just a single major locus. Using this model, I will analyse how effective this methodology is in identifying the causative locus when only a single major locus affects the value of the trait. In Chapter 3, I will extend the basic model developed in Chapter 2 to include selection at many loci. Using this extended model, I will analyse how effective this methodology is in identifying the causative loci for quantitative traits. I will also look at what effect epistasis has on marker frequency in the selected population. In Chapter 4, I will outline how to statistically analyse marker frequencies in the selected population using the model developed in Chapter 2.

Chapter 2: Mendelian Traits

Abstract

In this chapter the basic theoretical framework will be developed for the strategy of picking out the extreme individuals in pooled asexual cross progeny, by selecting for the trait over many generations. This chapter will concentrate on Mendelian traits. I will derive the distribution of marker frequency in a selected pool as a result of selection at just a single major locus. I will show that spurious changes in marker allele frequency in the selected population arise due the finite initial population size. Using the model developed I show that the initial population size should be at least in the low to mid hundreds to avoid spurious changes in marker frequency in the selected population

2.1 Model

A cross is made between two haploid lines that differ in trait value. This cross results in *N* haploid recombinant progeny, each containing a random assortment of marker alleles from the parental lines, with each marker having an expected frequency of 0.5. In this chapter I concentrate on the simplest situation where the variation in phenotype between the two lines is due to just one major locus. A fitness advantage is assigned to the recombinants that contain the positive allele (ie. the allele that increases the value of the trait), and so the initial population consists of two fitness classes. This recombinant population is then selected for the trait over many generations. As this population is asexual, no further recombination takes place during this multi-generation selection phase. It is assumed that selection is applied for long enough so that only recombinants originating from the fitter class remain in the final population. Therefore, the positive allele should be fixed in the selected population, and because there is only one round of

recombination, markers in a large region around the selected locus should also be at a higher frequency. The frequencies of markers in all other regions of the genome are expected to remain unchanged. So, from this model we are interested in analysing the frequency of all markers in the selected population, and the stochasticity that arises in this frequency due to finite population size.

2.2 Deterministic Expectation

If selection is continued until the fitter class of recombinants fix in the population, then the selected allele will be at frequency 1. The expected frequency of all other markers in the selected population would be equal to the probability that the marker in question was on the same genotype as the selected allele in the initial population. For the positive markers (fitter parental markers), this probability would simply be 1-*r*, and for the negative markers (less fit parental markers) it would be just *r*, where *r* is the probability of recombination between the selected allele and the marker in question.

2.3 Stochastics

With an infinite number of recombinants, the marker frequency will approach the deterministic expectation, but finite numbers will lead to variation around this expectation. In the extreme, suppose there was just one recombinant with the positive allele in the initial population. The typical marker composition of this recombinant will look like one of those given in Figure 2.1(a). As this single recombinant is the fittest in the initial population, selection (if applied for long enough) will pick out only its descendants. Therefore, all recombinants in the selected population will have exactly the same marker composition. Hence, the final marker frequencies will look like those in Figure 2.1(b), where a marker is either fixed or not present at all. With more than one initial recombinant with the positive allele present in the initial population, there will be

initially much more diversity in the marker composition, but this diversity may not be reflected in the final population. For example, suppose there were ten initial recombinants with the positive allele, each with a different marker composition. Again, selection will pick out only the descendants originating from these ten initial recombinants. However, the number of descendants that each recombinant actually leaves may be highly random. One may leave no descendants in the final population, while another may leave hundreds. Consequently, some markers will be over represented in the selected population, which can be seen from Figure 1(c) results in a very random pattern of marker frequency. This randomness is reduced by increasing the number of recombinants with the positive allele in the initial population. This results in a more balanced representation of all markers in the selected population. It can be seen from Figure 1(d) that with this increase in the number of recombinants with the positive allele in the initial population, the marker frequencies approach the deterministic expectation, enabling much easier identification of the selected locus. So, in order to evaluate how much stochasticity in marker frequency that would be expected for a certain initial population size, I will next derive the distribution of marker frequency in the selected population. From this, it is possible to calculate how large the initial population size needs to be in order to avoid spurious changes in marker frequency, and also work out the probability of getting false positives when we do have large stochasticity in frequency.

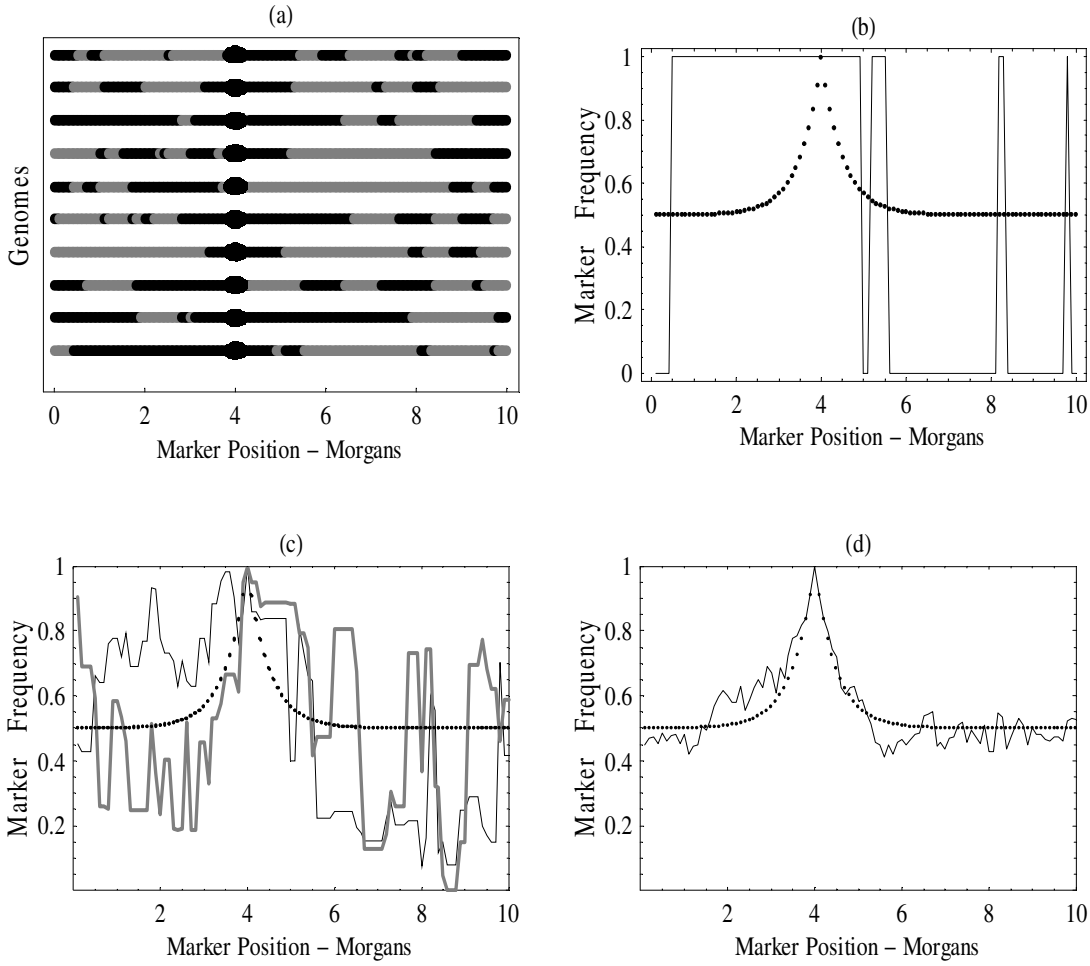


Figure 2.1 – (a) Each line represents the typical marker composition of a single recombinant with a selected allele at position 4 on the genome represented by a circle. The black parts represent the fitter parental markers (positive markers) and the gray parts represent the less fit parental markers (negative markers). (b) Plot of the positive marker frequency in the selected population when there is just a single recombinant (the first genome in (a)) in the initial population. (c) The black and gray curves show two replicates of the positive marker frequencies in the selected population when all ten recombinants from the first graph are present in the initial population. It can be seen that the two replicates do not give the same frequencies. This reflects the random number of descendents each recombinant left in each replicate. (d) This shows the frequency of the positive markers in the selected population when there are 100 recombinants with the positive allele in the initial population. In graphs (b), (c) and (d) the dotted curve represents the deterministic expectation for the positive marker frequency, which is $1-r$ where r is calculated from the Haldane map function $r = \frac{1}{2}(1-e^{-2x})$, and x is the map distance between the marker and selected locus.

2.4 Branching Process

To derive the distribution of marker frequency, the distribution of the number of descendants originating from a single recombinant needs to be obtained. This can be modelled as a branching process. That is, at each generation each selected recombinant leaves a number of offspring ξ, with mean μ and variance σ^2. This process can be modelled by the probability generating function $f(z) = \sum_0^\infty P_k z^k$, where P_k is the probability that $\xi = k$. This represents the offspring distribution of a single recombinant for a single generation. This can be extended to get the offspring distribution after t generations by t iterations of $f(z)$. That is $f_t(z) = f(f(\ldots(f(z))\ldots))$. So, if we let X denote the number of descendents originating from a single recombinant after t generations, we have that X has distribution $f_t(z)$. Obtaining probabilities from $f_t(z)$, however, can be computationally intensive, so instead just the moments of X will be outlined. From the properties of generating functions we have that the mean $E(X)$ and variance $Var(X)$ of the number of descendents originating from a single recombinant after t generations is given by (2.1) and (2.2) (Jagers, 1975)

$$E(X) = \mu^t \tag{2.1}$$

$$Var(X) = \sigma^2 \mu^{t-1}(\mu^t - 1)(\mu - 1)^{-1} \tag{2.2}$$

2.5 Moments

Using (2.1) and (2.2) it is possible to obtain the mean, variance and covariance of the number of copies of each marker in the selected population. Consider an initial population of size N and a marker m, and define S_m as the number of copies of that marker in the selected population. We have that $S_m = \sum_{i=0}^{n} X_i$ where n is a random variable representing the initial number of recombinants that had marker m. As we are assuming in the model that only the fitter class of recombinants survive in the selected population, we have that n is a binomially distributed random variable with

expectation $E(n) = 0.5NP_m$, where P_m is the probability that marker m is on the fittest genotype. Therefore, the expected number of copies of a marker m, $E(S_m)$, and variance $Var(S_m)$ in the selected population is given by (2.3) and (2.4).

$$E(S_m) = E(E(S_m \mid n)) = E(n)E(X) \tag{2.3}$$

$$\begin{aligned} Var(S_m) &= E(Var(S_m \mid n)) + Var(E(S_m \mid n)) \\ &= E(n)Var(X) + Var(n)E(X)^2 \end{aligned} \tag{2.4}$$

Given two markers m_1 and m_2, the covariance, $Cov(S_{m_1}, S_{m_2})$, between the number of copies of each marker in the selected population is given by (2.5), where $P_{m_1 m_2}$ is the probability that both markers m_1 and m_2 are on the fittest genotype.

$$Cov(S_{m_1}, S_{m_2}) = NE(X)^2(P_{m_1 m_2} - P_{m_1}P_{m_2}) + NVar(X)P_{m_1 m_2} \tag{2.5}$$

For the moments of marker frequency it is difficult to get exact expressions, so using (2.3), (2.4) and (2.5) we can get an approximation for the mean, variance and covariance in marker frequency in the selected population. Let $F_m = S_m / S_t$ be the frequency of marker m, where S_t is the total number of recombinants in the selected population. If we expand F_m as a Taylor series, we get (2.6) and (2.7) as an approximation for the mean and variance in marker frequency in the selected population. To derive the covariance in frequency, $Cov(F_{m_1}, F_{m_2}) = E(F_{m_1}F_{m_2}) - E(F_{m_1})E(F_{m_2})$, we can expand $E(F_{m_1}F_{m_2}) = (S_{m_1}S_{m_2}) / (S_t)^2$ as a Taylor series and get (2.8) as an approximation for the covariance in frequency between markers m_1 and m_2.

$$E(F_m) \approx \frac{E(S_m)}{E(S_t)} + \frac{Var(S_t)E(S_m)}{E(S_t)^3} - \frac{Cov(S_m, S_t)}{E(S_t)^2} \tag{2.6}$$

$$Var(F_m) \approx \frac{Var(S_m)}{E(S_t)^2} + \frac{E(S_m)^2 Var(S_t)}{E(S_t)^4} - \frac{2E(S_m)Cov(S_m, S_t)}{E(S_t)^3} \tag{2.7}$$

$$Cov(F_{m_1}, F_{m_2}) \approx \frac{E(S_{m_1})E(S_{m_2})}{E(S_t)^2} + \frac{Cov(S_{m_1}, S_{m_2})}{E(S_t)^2} - \frac{2E(S_{m_1})Cov(S_{m_2}, S_t)}{E(S_t)^3} - \frac{2E(S_{m_2})Cov(S_{m_1}, S_t)}{E(S_t)^3} + \frac{3E(S_{m_1})E(S_{m_2})Var(S_t)}{E(S_t)^4} - E(F_{m_1})E(F_{m_2}) \quad \textbf{(2.8)}$$

Again, since the model assumes that only recombinants from the fittest class survive in the selected population, some simplifications to the above calculations can be made. Given that only one fitness class survives we have that $E(S_m) = P_m E(S_t)$ and $Cov(S_m, S_t) = P_m Var(S_t)$, where $E(S_t)$ and $Var(S_t)$ can be calculated using (2.3) and (2.4) where n now is a binomial random variable with expectation $0.5N$. Substituting these into (2.6), (2.7) and (2.8) we get (2.6a) as the expectation of frequency, which is just the same as the deterministic expectation, and (2.7a) and (2.8a) as the variance and covariance in frequency.

$$E(F_m) = P_m \quad \textbf{(2.6a)}$$

$$Var(F_m) = 2P_m(1-P_m)\left(1 + \frac{Var(X)}{E(X)^2}\right)\frac{1}{N} \quad \textbf{(2.7a)}$$

$$Cov(F_{m_1}, F_{m_2}) = 2(P_{m_1 m_2} - P_{m_1}P_{m_2})\left(1 + \frac{Var(X)}{E(X)^2}\right)\frac{1}{N} \quad \textbf{(2.8a)}$$

2.6 Diffusion Approximation

Although expressions for the moments of the number of copies of a marker and moments for the frequency of a marker have been obtained, in order to obtain a tractable expression for the distribution of these we need to use a diffusion approximation. Diffusion theory predicts (Feller, 1951), that starting with n_0 copies, after a long time, given that they survive, the numbers will increase as $n_0 x e^{st}$, where $0 < x < \infty$ is a

measure of the acceleration relative to the expectation $n_0 e^{st}$, and its distribution is given by

$$\phi(x) = \frac{2e^{-2n_0 sx} n_0 s I_1(4n_0 s\sqrt{x})}{(e^{2n_0 s} - 1)\sqrt{x}} \quad \textbf{(2.9)}$$

where $I_1(x)$ is the modified Bessel function and $s = log(\mu)$. For small $n_0 s$ (2.9) approximates to an exponential distribution. So, as an approximation we can try and use an exponential distribution for the distribution of numbers from a single recombinant. The expected value λ^{-1} for the exponential distribution would be the expected x of a single recombinant given that its descendents have survived in the selected population. We have that the probability of survival $P_S = 1 - f_t(0)$, and thus $\lambda^{-1} = P_S^{-1}$. So, therefore we get (2.10) as an approximation for the distribution of x

$$\varphi(x) = P_S e^{-P_S x} \quad \textbf{(2.10)}$$

It should be noted, however, that as (2.10) is an approximation derived from the diffusion result, which itself is an approximation of the general branching process, it is not expected it will work well in all situations. Figure 2.2 shows the goodness of fit of (2.9) and (2.10) for simulated data. It can be seen that both work well for weak selection but decline in goodness of fit for strong selection. So, in the following section I will use (2.10) to derive the distribution of marker frequency for situations when fitness is not too high, but as I shall show later, for large fitness we can in most cases use a normal approximation for the distribution of frequency using the moment calculations (2.6a) – (2.8a).

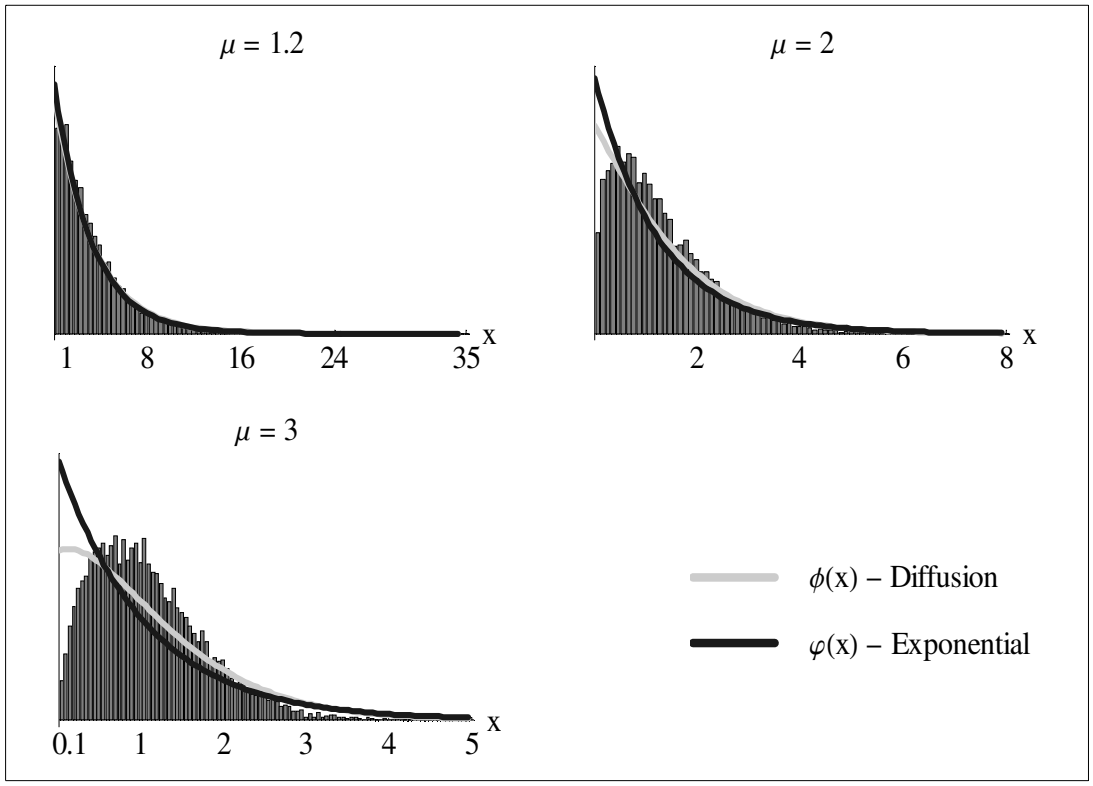


Figure 2.2 – Distribution of the relative numbers from of a single recombinant given that its descendants have survived in the selected population. The diffusion curve represents (2.9) with parameters $n_0 = 1$ and $s = log(\mu)$, and the exponential curve represents (2.10). The number of generations of growth were $t = \{20, 10, 10\}$ for $\mu = \{1.2, 2, 3\}$. The offspring distribution per generation was Poisson.

2.7 Distribution of Marker Frequency

I will assume that the distribution of the number of descendents from a single recombinant, given that its descendents have survived in the selected population, is an exponential distribution with expectation $E(X)P_S^{-1}$. Now consider an initial population of size N and a single positive marker (ie. a marker from the fitter parental strain) m^+ a recombination rate r away from the selected locus. We have that the number of copies of m^+ in the selected population is given by $S_m^+ = \sum_{i=0}^{n_1} X_i$, where each of X_i is exponentially distributed and n_1 is a binomially distributed random variable with

expectation $E(n_1) = \frac{1}{2} NP_S(1-r)$. Thus S_m^+ is distributed as $\Gamma(n_1, E(X)P_S^{-1})$, where Γ represents a Gamma distribution (ie. a sum of exponential distributions). So, the frequency of m^+ in the selected population would be defined as $S_m^+ / (S_m^- + S_m^+)$ where S_m^- is the number of negative markers at that locus in the selected population, which has distribution $\Gamma(n_2, E(X)P_S^{-1})$ where $E(n_2) = \frac{1}{2} NP_S r$. Hence, the distribution of marker frequency is a Beta distribution B(n_1, n_2). Averaging over n_1 and n_2, we get (2.11) as the probability density function for a positive marker frequency u, where $p_1 = \frac{1}{2} P_S(1-r)$ and $p_2 = \frac{1}{2} P_S r$.

$$f(u) = \sum_{n_1=1}^{N} \sum_{n_2=1}^{N} \frac{N!}{n_1!(N-n_1)!} \frac{N!}{n_2!(N-n_2)!} {p_1}^{n_1} (1-p_1)^{N-n_1} {p_2}^{n_2} (1-p_2)^{N-n_2} \frac{\Gamma(n_1+n_2)}{\Gamma(n_1)\Gamma(n_2)} u^{n_1-1} (1-u)^{n_2-1} \quad \textbf{(2.11)}$$

It should be noted that as the Beta distribution is only defined for $n_1, n_2 > 0$, $f(u)$ does not take into account the case where there are zero copies of a particular marker at the locus (ie. $n_1 = 0$ or $n_2 = 0$). This results in the density function $f(u)$ excluding the probability that a marker is fixed or lost in the selected population. Therefore the true density function is given by $f(u) + P(u=0) + P(u=1)$, where $P(u=0)$ is the probability that the marker is lost, and $P(u=1)$ is the probability that the marker is fixed. If we again focus on a positive marker m^+, we have that $P(u=1) = (1-(1-p_1)^N)(1-p_2)^N$, where $(1-(1-p_1)^N)$ is the probability that at least one recombinant with marker m^+ survives in the selected population, and $(1-p_2)^N$ is the probability that no recombinants with the negative marker at that locus survives in the selected population. Similarly $P(u=0) = (1-(1-p_2)^N)(1-p_1)^N$. It should be noted that the inclusion of these two probabilities is only really needed in the cases where the initial population size is very small or when a marker is extremely close to the selected locus, as the probability of a marker being fixed or lost in other situations is negligible.

Figure 2.3 illustrates the goodness of fit of this approximation for various different parameters. We see, as expected, (2.11) works well for small μ but goodness of fit declines as μ gets larger. For large μ, however, assuming N is not too small, we can approximate the distribution of frequency by using a normal distribution with mean and variance given by (2.6a) and (2.7a). It can be seen from Figure 2.3(c) and 2.3(d) that the normal distribution provides a good approximation when the initial population is not too small.

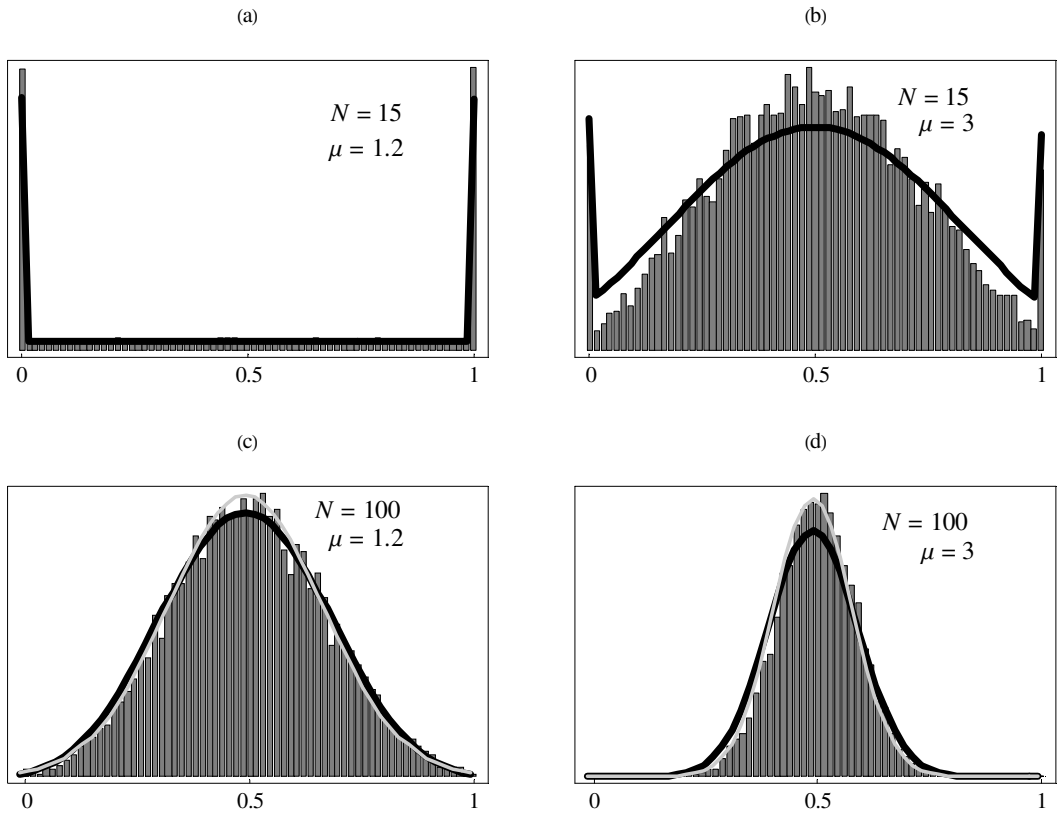


Figure 2.3 – Distribution of frequency for unlinked markers for a small initial population size of $N = 15$ and a larger initial population size of $N = 100$. For each of the initial population sizes, the distribution of frequency is plotted for small fitness $\mu = 1.2$ and large fitness $\mu = 3$. The black curve represents (2.11) while the grey curve in (c) and (d) is a normal approximation using (2.6a) and (2.7a). The number of generations of selection was 20 for $\mu = 1.2$ and 10 for $\mu = 3$. The offspring distribution was Poisson.

2.8 Effective Initial Population Size

Using the moment calculations it is possible to work out how large the initial population size N should be in order to avoid spurious changes in marker frequency. As seen in Figure 2.1 the larger N is, the less variation in frequency in the selected population. However, it can also be seen from Figure 2.3 that even though the same initial population size can be present in two of the same experiments, the distribution of marker frequency can be very different. In Figure 2.3(a) and 2.3(b), both simulations show large variation in frequency due to having only a small initial population size of $N = 15$. Figure 2.3(a), however, shows far more variation than Figure 2.3(b). This discrepancy is due to the variation in the number of descendants each initial recombinant leaves in the selected population. The majority of this variation in the number of descendants can be attributed to the differences in the probability of survival of the initial recombinants in the two examples. That is, not all of the 15 recombinants in the initial population have survived and left descendants in the selected population. Only a certain portion of the initial population have actually contributed towards the final frequency. This subset of the initial population that actually leave descendents in the selected population is what I will refer to as the effective initial population size N^*. Since, it is assumed in this model that only the fittest genotype remains in the selected population, this effective initial population size N^* can be defined as the initial proportion of recombinants within this fitter class that leave descendents in the selected population. As a result, N^* is a binomially distributed random variable with $E(N^*) = 0.5NP_S$. The larger N^* is, the less the variation in marker frequency. For instance, in Figure 2.3(a) the probability of survival $P_S = 0.32$, and hence $E(N^*) = 2.38$, while in Figure 2.3(b) $P_S = 0.94$ and $E(N^*) = 7.05$. So, although both examples started off with 15 unique recombinant genotypes, on average only about 2 unique genotypes are represented in the selected population in one example, whereas on average 7 unique genotypes are represented in the selected population in the other. So, this reduction in the effective initial population size led to a lot more variation in frequency in the example in Figure 2.3(a). The same

explanation is responsible for the differences in marker distribution in Figures 2.3(c) and 2.3(d). Hence, when determining how large the initial population size N should be, one needs to take into account the probability of survival. In general, when the mean offspring per generation is small, the probability of survival would be quite low and a much larger N would be needed to ensure enough unique genotypes survive in the selected population. This can be seen in Figure 2.4. Using a Poisson distribution of offspring, Figure 2.4 plots the variance in frequency in the selected population (using (2.7a)) against N for various different fitnesses. It can be seen that, as expected, for small N there is a lot more variation, and for small μ the variance is even larger due to the smaller N^*. It can also be seen that having an initial population size at least in the mid hundreds ensures only small variation in marker frequency in the selected population.

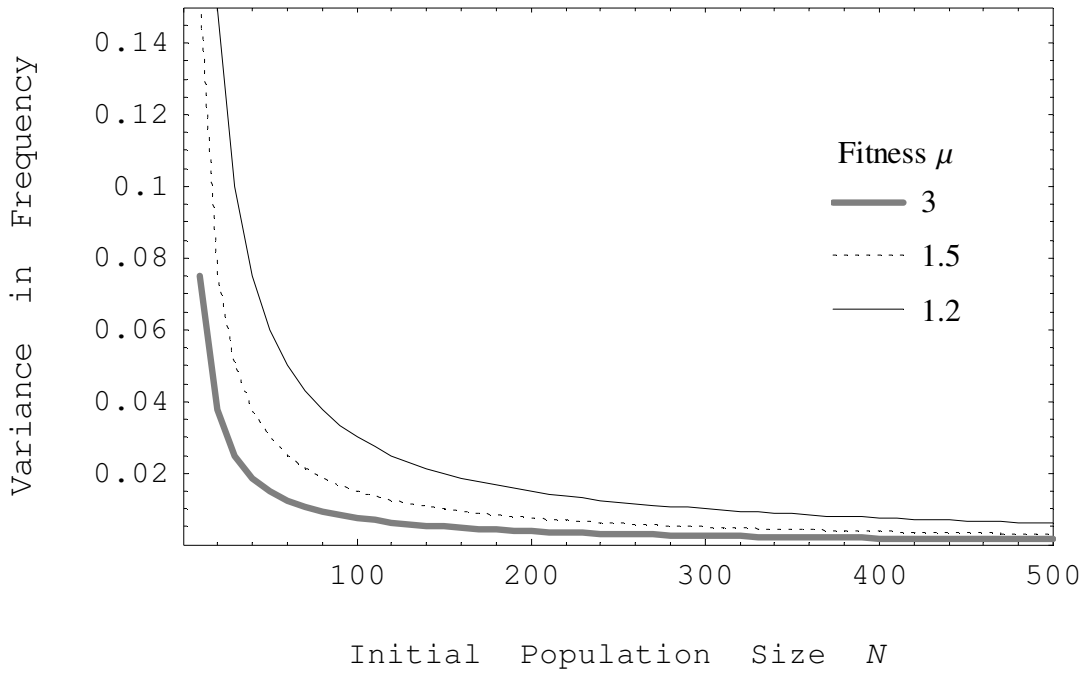


Figure 2.4 – The variance in frequency in the selected population for unlinked markers. The variance was calculated from $2 P_m (1 - P_m) (\mu^2 + \sigma^2 - \mu) (N (\mu - 1) \mu)^{-1}$ (ie. limit of (2.7a) as $t \to \infty$) where $P_m = 0.5$ and $\mu = \{1.2, 1.5, 3\}$. The offspring distribution used was a Poisson distribution and thus $\sigma^2 = \mu$.

2.9 False Positives

To get an idea of how this variation in marker frequency affects the mapping ability, we can calculate the number of false positives we would get, when we try to identify markers linked to the selected locus. For instance, suppose we wanted to do an initial genome scan to see which chromosome the selected allele lies on. The deterministic expectation predicts that the closer a particular marker is to the selected locus the more extreme the frequency of that marker becomes. Hence, identifying the marker with the highest (positive markers) or lowest (negative markers) frequency should reveal, at a minimum, which chromosome the selected allele lies on. Finite population sizes, however, may lead to more extreme marker frequency on other chromosomes. So, for various initial population sizes, what is the probability that the most extreme marker frequency is the marker that is linked to the selected locus? If we look at the positive markers we are interested in finding the maximum marker frequency. In this case, we can define a false positive as a marker in unlinked regions that has a frequency greater than the marker that is closest to the selected locus. Hence, we need to evaluate $P(u_{null} < u_{linked})$ where u_{null} is the maximum frequency in unlinked (or null) regions, and u_{linked} is the frequency of the marker closest to the selected locus. To evaluate this probability, I will assume that there are c chromosomes of equal length l Morgans, and assume each chromosome has a total of τ markers at equally spaced intervals $d = l/(\tau - 1)$. For simplicity, I will also assume that the selected allele is positioned in the middle of two markers resulting in the distance between the closest marker and the selected allele being $d/2$. Now, in order to evaluate the distributions for u_{null} and u_{linked}, I will use the normal approximations using moment calculations (2.6a) – (2.8a). So, let $f_N(u_{linked})$ be the normal approximation for the probability density of u_{linked}, and let $P(u_{linked} = 1)$ be the probability that u_{linked} is fixed in the selected population. For u_{null}, the distribution of the maximum frequency from the set of markers in unlinked regions is needed. We need to use a multivariate normal distribution for this probability as the frequencies of markers on the same chromosome can be correlated. So, for any given

value of u_{linked}, an approximate probability that the maximum frequency in unlinked regions is less than u_{linked}, is given by $P(u_{null} < u_{linked}) = F_{CMVN}(\mathbf{u})^{c-1}$, where $F_{CMVN}(\mathbf{u})$ is the cumulative multivariate normal distribution, and $\mathbf{u}$ is a vector of length τ with all elements equal to u_{linked}. Integrating over all possible values of u_{linked} we get (2.12) as an approximation for the probability of not getting a false positive.

$$P(u_{null} < u_{linked}) = \int_0^1 F_{CMVN}(\mathbf{u})^{c-1} f_N(u_{linked}) du_{linked} + P(u_{linked} = 1) \quad \textbf{(2.12)}$$

Figure 2.5 shows how well (2.12) works against simulation results. The solid curves are the theoretical results using (2.12) and the dashed curves are the corresponding results from simulations. The curves plot the probability of getting a false positive for increasing effective initial population size. In the example, there are $c = 20$ chromosomes each of length $l = 1$ Morgan. The false positive probabilities were calculated when there were $\tau = 3$ and $\tau = 5$ markers per chromosome. It can be seen that the approximation (2.12) slightly overestimates the number of false positives. This is mainly due to the normal approximation for u_{linked}. As a marker becomes closer to the selected locus, the less it follows a normal distribution. As a result, the false positive rate is overestimated. For extremely small initial population sizes, (2.12) would not provide a good approximation for the number of false positives, as the marker frequencies can no longer be approximated by a normal distribution. In general, however, it can be seen from Figure 2.5, that the false positive rate is reduced, as expected, when the variation in marker frequency is reduced with the increase in the effective initial population size. With the smaller effective initial population sizes, an increase in the marker density is needed to reduce the number of false positives. It should also be noted that with extremely small initial population sizes (ie. effective initial population size less than 15), the probability of fixation of a marker in unlinked regions is greater than zero, and as a result the false positive rate may always remain high no matter how densely the markers are spaced.

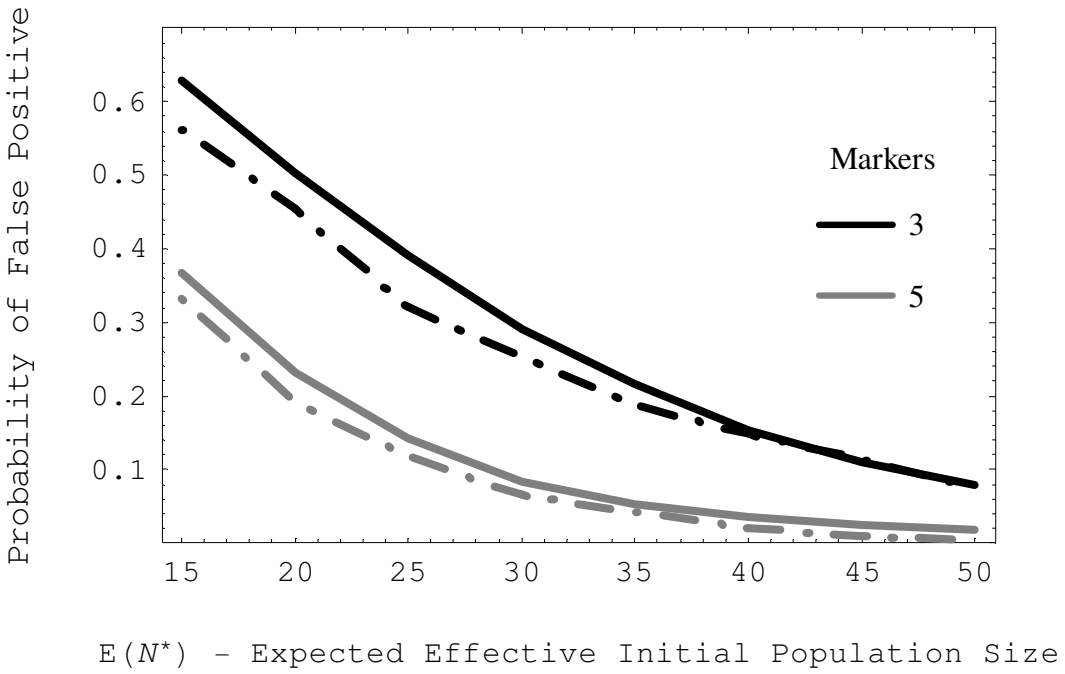


Figure 2.5 – The probability of getting a false positive plotted against the expected effective initial population size $E(N^*)$. The solid curves are the theoretical predictions using (2.12) (ie. $1-P(u_{null} < u_{linked})$) and the two dashed curves are simulation results. The parameters that were used was c=20 chromosomes each of length l=1 Morgan. The number of markers τ on each chromosome was τ=3 and τ=5. The black curves are results when τ=3 and the gray curves are the results when τ=5. The number of generations of selection was 10 and overall fitness of selected allele was 3.

2.10 Discussion

The aim is to locate alleles that influence a trait by examining changes in marker allele frequency in selected progeny. The extreme progeny are selected by multiple generations of asexual reproduction and selection. I have shown that when just a single major locus affects the trait, the accuracy in identifying markers linked to the causative allele depends on the variance in marker frequency in the selected population, which in turn depends on the effective initial population size N^*. This effective initial population size N^* was defined as the number of unique recombinant genotypes in the initial

population that actually leave descendants in the selected population. The larger N^* is, the less variation in marker frequency in unlinked regions in the selected population. From Figure 2.4 it was shown that having an initial population size in the mid hundreds ensures that the marker frequencies in the selected population approach the deterministic expectation. In such a situation, there is only a small probability of spurious changes in marker frequency in unlinked regions, and thus it should be relatively easy to detect the general location of the selected locus.

The ease of detection will also depend on the marker density. Having a very dense map of markers will ensure that a marker is close enough to the selected locus, so that its frequency is the most extreme in genome, making identification of the location of the selected locus easier. How dense the markers need to be to achieve this will mainly be determined by the effective initial population size, and also by the length and number of chromosomes. From the example in Figure 2.5, where there were 20 chromosomes each of length 1 Morgan, relatively few markers were needed per chromosome to achieve a low false positive rate, as long the effective initial population size was not too small. This perhaps explains why this technique has been successful in mapping simple traits. That is, achieving an initial population size of a few hundred recombinants, particularly in microbial populations where thus technique has mostly been used, is very achievable, making this a relatively efficient method for mapping simple Mendelian traits.

Chapter 3: Quantitative Traits

Abstract

The aim of this chapter is to analyse the behaviour of marker frequency when more than one locus affects the value of the trait. In the previous chapter, when the initial population consisted of just two fitness classes, it was assumed that only the fitter class of recombinants survived in the population. When many loci influence the trait, I will show that selecting out only the fittest class of recombinants may not be desirable or even possible. It may not be desirable as only selecting the fittest class of recombinants may lead to large stochasticity in marker frequency in unlinked regions. This is due to the reduction in the effective initial population size as selection is applied. Also, when a large number of loci influence a trait, it may not be possible to select the fittest possible class of recombinants, as it may not be present in the initial population due to low probability of being produced at meiosis. I will show that this can result in different fitness classes establishing in the population in different replicates, which may lead to more variability in marker frequency at selected loci. Interaction between selected loci can also occur when many loci influence the trait. I will show that when interaction occurs between many selected loci, appreciable changes in marker frequency at selected loci will on average take longer to happen than in an additive model. Consequently the presence of epistasis hinders the ability to identify selected loci. I will also show that it is possible to detect epistasis from the marker frequency by identifying reversals in the direction of marker frequency change.

3.1 Introduction

The aim is to locate alleles that influence a trait by examining changes in marker allele frequency in selected progeny. The extreme progeny are selected by multiple generations of asexual reproduction and selection. In the previous chapter the variation in trait value between the two parental lines in a cross was assumed to be caused by a single genetic locus. This resulted in two fitness classes in the cross. Selection was then applied for the trait in this cross progeny until only the fittest recombinants survived. Marker frequencies were then evaluated in these fitter individuals. In this chapter, the aim is to apply this same strategy to quantitative traits. I will show, however, that when the trait is influenced by many loci, selecting out the fittest possible class of recombinants may not be possible or desirable.

To illustrate, suppose there are l loci that affect the value of the trait. There could be alleles that increase or decrease the value of the trait on either of the parental lines. A cross between these parental lines could now result in a possible 2^l genotypes in the initial recombinant population. If selection is continued for long enough, the population will fix for one of these genotypes. Now, unlike the one selected locus case, in any one replicate the genotype that fixes may not be the fittest possible. This is because with multiple selected loci, the probability that the fittest possible genotype is produced at meiosis may be quite small. For example, if ten unlinked loci influence the value of the trait, then the probability that the fittest genotype is produced at meiosis is just 2^{-10}. Hence, unless the initial population size is quite large, the fittest genotype would most likely not be present. Even if it were present in the initial population, it would probably be there at low numbers, and as a result it could easily undergo stochastic loss in the initial few generations. Therefore with large l, in any one replicate, the genotype that fixes will be the fittest genotype that has survived the initial few generations. This may not be the fittest possible genotype but will be one of the genotypes in the upper tail of the fitness distribution.

There are two main issues that arise with this situation. Firstly, since different fitness classes can now potentially establish in the population, I will show that this may result in more variability in frequency at selected loci. Secondly, as each fitness class would most likely have been at low numbers in the initial population, selecting until a genotype fixes will lead to large variability in frequency in unlinked regions. As a result, when a large number of loci influence the trait, finding optimal selection times would be appropriate. Therefore, in this chapter I will attempt to find optimal selection times, but I will show it is a difficult process. As a result I will only give very general guidelines on how long selection should be applied. Finally, the role of epistasis will be examined, and it will be shown that interaction between large numbers of selected loci will generally increase the difficulty in identifying selected alleles, as appreciable changes in marker frequency at selected loci will on average take longer to happen than in an additive model. Lastly, I will show that in some models of epistasis it is possible to detect interaction between selected loci from marker frequencies by identifying reversals in the direction of marker frequency change.

3.2 Model

A cross is made between two haploid lines that differ in trait value. This cross results in *N* haploid recombinant progeny each containing a random assortment of marker alleles from the parental lines. Each marker is expected to have an initial frequency of 0.5. In this initial population each allele is assigned a relative fitness of $1+s$, where $s=0$ for neutral alleles, $s>0$ for alleles that increase the value of the trait and $-1 \le s < 0$ for alleles that decrease the value of the trait. It is assumed that there are $l \ge 1$ loci responsible for the differences in trait value between the two lines. Therefore there could be a possible 2^l genotypes in the initial population. Each genotype is given a fitness based on a multiplicative model. That is $w_j = \prod_i (1+s_i)$, where w_j refers to the fitness of the j^{th} genotype, and the s_i refers to the selection coefficient of the i^{th} allele

on each genotype. This initial recombinant population is then selected for the trait over many generations. As this population is asexual no further recombination takes place during this multi generation selection phase. Marker frequencies are then observed in the selected population at some generation *t*. From this model we are interested in calculating the frequency of all markers in the selected population, and determining the stochasticity that arises in this frequency due to finite population size.

3.3 Deterministic Expectation

Suppose selection was continued for long enough for a genotype *G* to fix in the selected population. The expected frequency, $E(F_m)$, of a marker *m* in the selected population is given by $E(F_m) = P_G^* / P_G$, where P_G is the probability that the genotype *G* that has fixed was produced at meiosis, and P_G^* is the probability that marker *m* was also on genotype *G* at meiosis. This probability simplifies to just the recombination probabilities between the marker and the immediate flanking selected alleles. That is, $E(F_m) = P_G^* / P_G = (P_L * P_R) / P_{LR}$, where P_L is the probability marker *m* and the selected allele immediately to the left of the marker are on the same genotype at meiosis, P_R is the probability marker *m* and the selected allele immediately to the right of the marker are on the same genotype at meiosis, and P_{LR} is the probability that both immediate flanking selected alleles are on the same genotype at meiosis. Using this Figure 3.1 shows a simple plot of the expected marker frequency when there are five selected loci. There are five clear peaks in the marker frequency with a large area around each selected locus showing an increase in frequency.

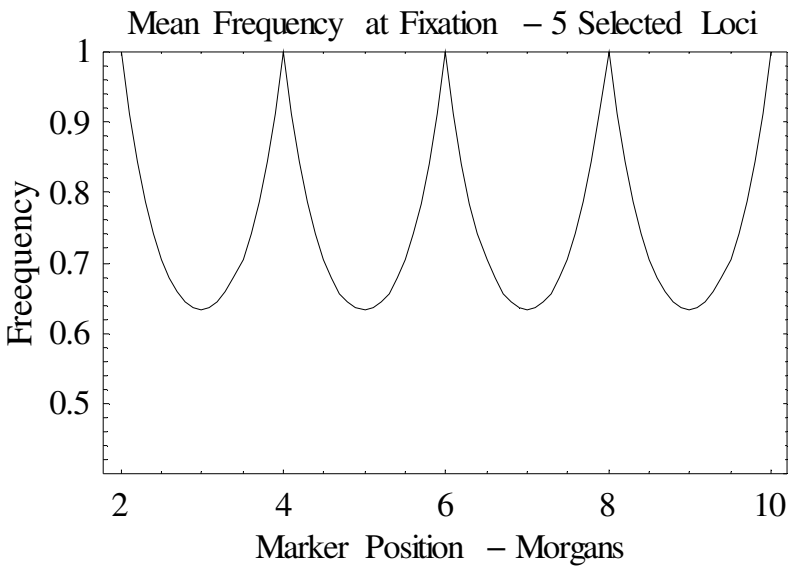


Figure 3.1 – Plot of the expected marker frequency at fixation when there are 5 selected loci at positions {2, 4, 6, 8, 10}.

3.4 Stochastics

Given that selection is continued until a genotype fixes, marker frequencies will approach the deterministic expectation shown in Figure 3.1 if an infinite initial population size was present. However, with finite population sizes, Figure 3.1 is only typical of what would be seen if selection was continued until a genotype fixes, and there was an extremely large initial population size present and/or marker frequencies were averaged across a large number of replicate experiments. In any one replicate experiment, however, the marker frequencies that are seen may display a very different pattern if selection was continued until a genotype fixes. This is because the longer selection is applied for, the more stochasicity that would be seen in marker frequency in unlinked regions. This is due to the reduction in the effective initial population size as selected is applied. That is, if a large number of loci influence the trait, then each fitness class will be at low numbers in the initial population. If selection is continued for long enough and a single fitness class fixes, then all recombinants in the selected population will have originated from just a few unique recombinant genotypes. As detailed in the

previous chapter, this would lead to an unbalanced representation of markers in the selected population, and consequently there will be large randomness in marker frequency in unlinked regions.

An example of this is shown in Figure 3.2. It shows the marker frequencies at various generations of selection, when there are five unlinked selected loci, one large effect locus and four small effect loci, and a relatively large initial population size of 200. The bar charts in Figure 3.2 represent the genotypic composition of the population at that particular generation. With five unlinked selected loci, there are $2^5 = 32$ possible genotypes, with each genotype having a probability $2^{-5} = 0.03125$ of being produced at meiosis. So, in the bar charts in Figure 3.2, each bar represents one of these 32 genotypes, with bar number 1 representing the least fit genotype and bar number 32 representing the fittest possible genotype. In the initial cross, it can be seen that most genotypes are equally represented in the population and markers frequencies are, as expected, around 0.5. After ten generations of selection, it can be seen that most genotypes are still present in the population, but the frequency of the genotypes in the upper half of the fitness distribution have increased. These genotypes in the upper half of the fitness distribution all have the large effect allele, and consequently it can be seen that the frequencies of markers around the large effect locus have increased. The frequencies of all other markers remain roughly the same. After thirty generations of selection it can be seen that the fitter genotypes are now starting to establish in the population, which results in an increase in frequency of the smaller effect alleles. It can also be seen that a lot of the genotypes in the lower half of the fitness distribution are at insignificant numbers or no longer present in the population. This results in a decrease in the effective initial population size. That is, after thirty generations of selection, the number of unique recombinant genotypes in the population has been reduced from 200 to 92. This results in slightly more variation in frequency in unlinked regions. After one hundred generations of selection, there are only six fitness classes present in the population, with the fittest (genotype 32) being the only one in substantial numbers, which results in the frequency of all the selected alleles nearing fixation. However, with

so few fitness classes remaining in the population, the effective initial population size has become very small. There are now only 27 unique recombinant genotypes in the population, with the vast majority of the population originating from just 6 unique recombinant genotypes. Consequently, many markers in unlinked regions are also at very low or high frequency.

So, it can be seen that in any one replicate, if selection is continued for a very long time, it may be very difficult to identify which of the peaks and valleys in marker frequency are truly selected alleles and which are from null regions, due to the very low effective initial population size. In order to avoid this, much larger initial population sizes would be needed so that enough numbers of the fitter genotypes are produced at meiosis. The moments in marker frequency can be used to get an idea of how much variation in marker frequency would be expected for a particular initial population size and selection time.

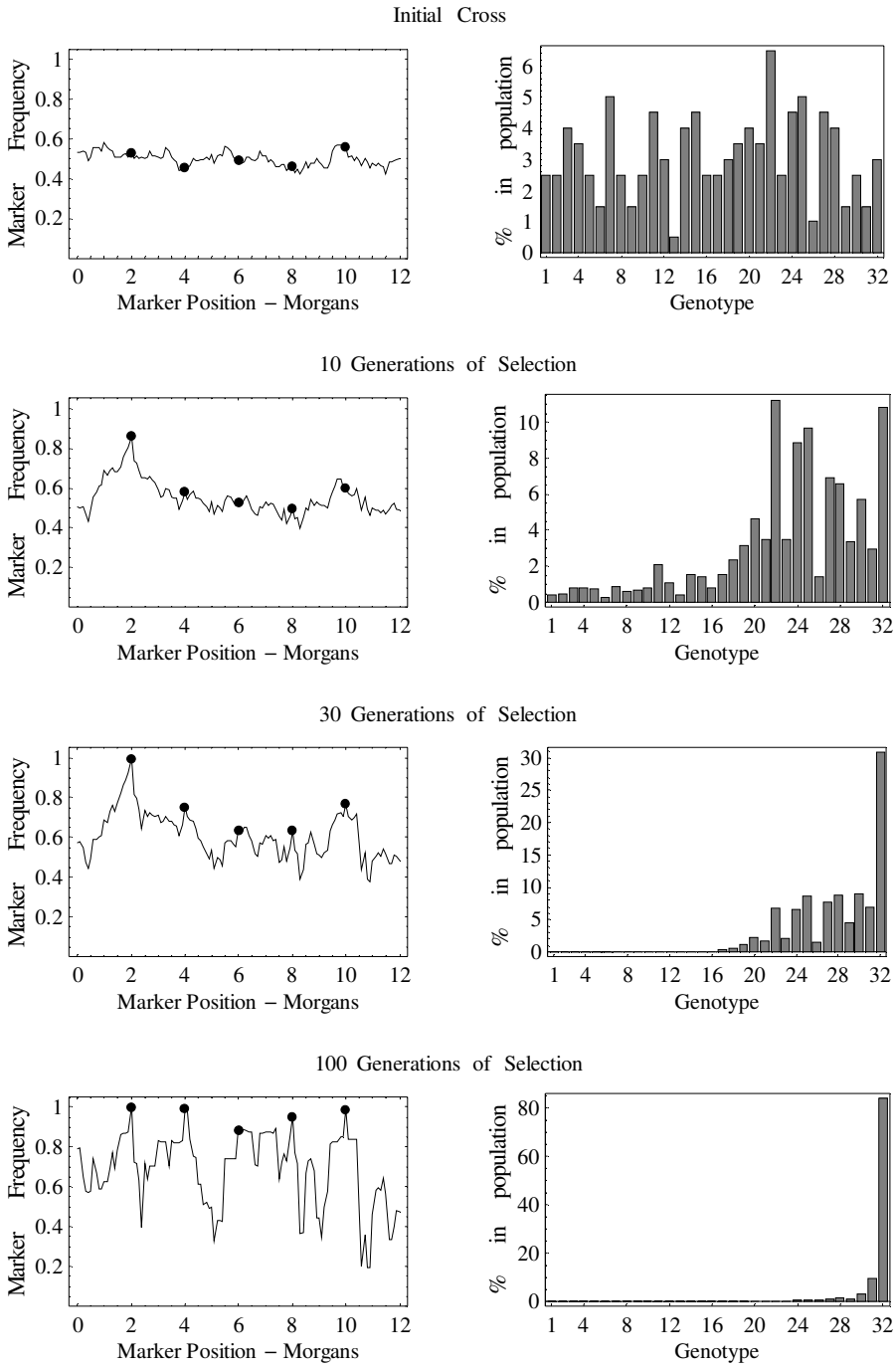


Figure 3.2 – This figure shows the marker frequencies and the genotypic composition of the population at various generations of selection when there are multiple selected loci. There are a total of five selected alleles at positions {2, 4, 6, 8, 10} (shown by the filled circles) with selection coefficients {0.2, 0.05, 0.01, 0.03, 0.04}. With five selected alleles there are 32 possible genotypes. The bar charts show the proportion of each of these 32 genotypes in the

population at that particular generation. Genotype number 1 refers to the least fit genotype (relative fitness of 1) and genotype 32 refers to the fittest possible genotype (relative fitness 1.36). The initial population size was 200.

3.5 Moments of Marker Frequency

The moments in marker frequency can be obtained by extending the single fitness class results which were derived in Chapter 2. In the single fitness class results expressions for the moments of marker frequency were obtained using Taylor series approximations. These Taylor series functions (2.6) – (2.8) will again be used for the moments in frequency in this chapter. To use these functions, the moments of the number of copies of each marker in the selected population were needed. So, given now there could be n fitness classes in the selected population, the number of copies S_m of a marker m can be defined as $S_m = \sum_{j=1}^{n} S_m^j$, where S_m^j is the number copies of marker m on the j^{th} fitness class. The moments of S_m is given by (3.1) – (3.3). Each element in (3.1) – (3.3) can be evaluated using the single fitness class results obtained in Chapter 2. It should also be noted that since these are only approximations they do not always provide accurate results. In particular, the Taylor series functions decline in accuracy as the number of individuals in each fitness class declines. Hence the moments in frequency may not provide accurate results if the number of individuals in each fitness class is quite small.

$$E(S_m) = E(\sum_{j=1}^{n} S_m^j) = \sum_{j=1}^{n} E(S_m^j) \quad \textbf{(3.1)}$$

$$Var(S_m) = Var(\sum_{j=1}^{n} S_m^j) = \sum_{j=1}^{n} \sum_{k=1}^{n} Cov(S_m^j, S_m^k) \quad \textbf{(3.2)}$$

$$Cov(S_{m_1}, S_{m_2}) = Cov(\sum_{j=1}^{n} S_{m_1}^j, \sum_{k=1}^{n} S_{m_2}^k) = \sum_{j=1}^{n} \sum_{k=1}^{n} Cov(S_{m_1}^j, S_{m_2}^k) \quad \textbf{(3.3)}$$

Using (3.1) – (3.3) and (2.7) Figure 3.3 plots the variance in marker frequency over time for the example shown in Figure 3.2. It plots the variance in frequency for the selected allele at position 4 which has an effect 0.05 (gray curve), the selected allele at position 6

which has an effect 0.01 (dashed curve), and finally the variance in frequency for an unlinked marker (black curve). It can be seen that for the unlinked marker that the variance in frequency, as expected, increases as selection is applied. For the selected alleles, the variance in frequency initially increases but then decrease as the selected alleles approach fixation, with the smaller effect allele taking a lot longer to reach fixation. So, assuming that the same genotype establishes in the population in each replicate, apart from markers very tightly linked to selected loci, the longer selection is applied for the more variation we would expect to see in marker frequency.

So, using (3.1) – (3.3) and (2.7) it is possible to get an idea of how large the initial population size should be in order to minimize the variation in frequency. However, if a very large number of loci influence the trait, then the initial population sizes that are needed may become prohibitively large. Hence, in this situation finding an optimal selection time would be more appropriate. Also, in some experiments selecting for a very long period of time may not be feasible, and finding an optimal selection time in these cases would also be useful. So, later in the chapter I will attempt to find optimal selection times, but before that I outline the behaviour of marker frequency when different genotypes establish in the population in different replicates. I will show that, unlike in Figure 3.3, when different genotypes establish in the population in different replicates, the variance in frequency of the smaller effect selected alleles and tightly linked markers will increase as selection is applied.

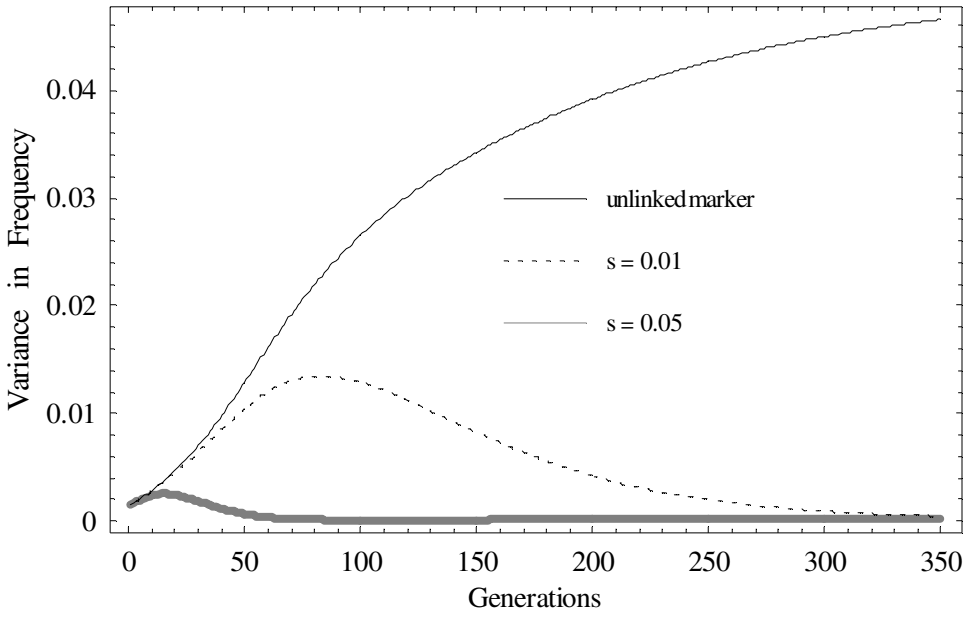


Figure 3.3 – Plot of the variance in marker frequency over time for three of the markers shown in the example in Figure 3.2. The gray curve represents the selected locus at position 4. The dotted curve represents the selected locus at position 6. The black represents a marker in an unlinked region.

3.6 Genotype Fixation Probability

In the example in Figure 3.2 the initial population size was large enough so that there was a high probability that the fittest possible genotype was always produced at meiosis and established in the population. If, however, the initial population size was much smaller, or a larger number of loci influenced the trait, then the probability that the fittest possible genotype was produced at meiosis and established in the population in each replicate would be much lower. There would be far more variability in which fitness class establishes in the population. This probability of genotype fixation can be evaluated as follows. The probability that a particular genotype G is produced at meiosis and establishes beyond the first few generations is given by $1-(1-P_G P_S)^N$ where P_G is the probability that genotype G is produced at meiosis, P_S is the probability of survival of that genotype, and N is the initial population size If we assume no crossover

interference between loci then $P_G = 0.5\prod_{i=1}^{l-1} P_i$, where P_i is simply the probability that the i^{th} consecutive pair of selected alleles are on the same genotype at meiosis. So, given that we have n fitness classes in the initial population, and selection is applied for long enough for a genotype to fix, the probability P_f^j that the j^{th} fitness class fixes is given by (3.4) (where $j = n$ is the fittest)

$$P_f^j = 1 - (1 - P_G^j P_S^j)^N \prod_{i=j+1}^{n} (1 - P_G^i P_S^i)^N \quad \textbf{(3.4)}$$

Using (3.4) Figure 3.4(a) illustrates the probability of fixation for the 32 genotypes in the example in Figure 3.2 when the initial population size was 200 (black filled bars) and 50 (gray filled bars). It can be seen that when the initial population size is 200, the fittest genotype will always establish in the population. When the initial population size is 50, there is slightly more variability in which genotype fixes, but still a high probability the fittest possible genotype will establish. With a larger number of selected loci and/or a smaller initial population size, there would be a lot more variability in the fixation probabilities. Figure 3.4(b) shows an example of this. It shows the fixation probabilities of the top 20 genotypes when there are 7 selected loci when the initial population size was 200 (black filled bars) and 50 (gray filled bars). It can be seen that when the initial population size is 50, there is only a 40% probability that the fittest possible genotype establishes, and hence in repeated experiments we would expect a lot of variability in which genotype establishes in population.

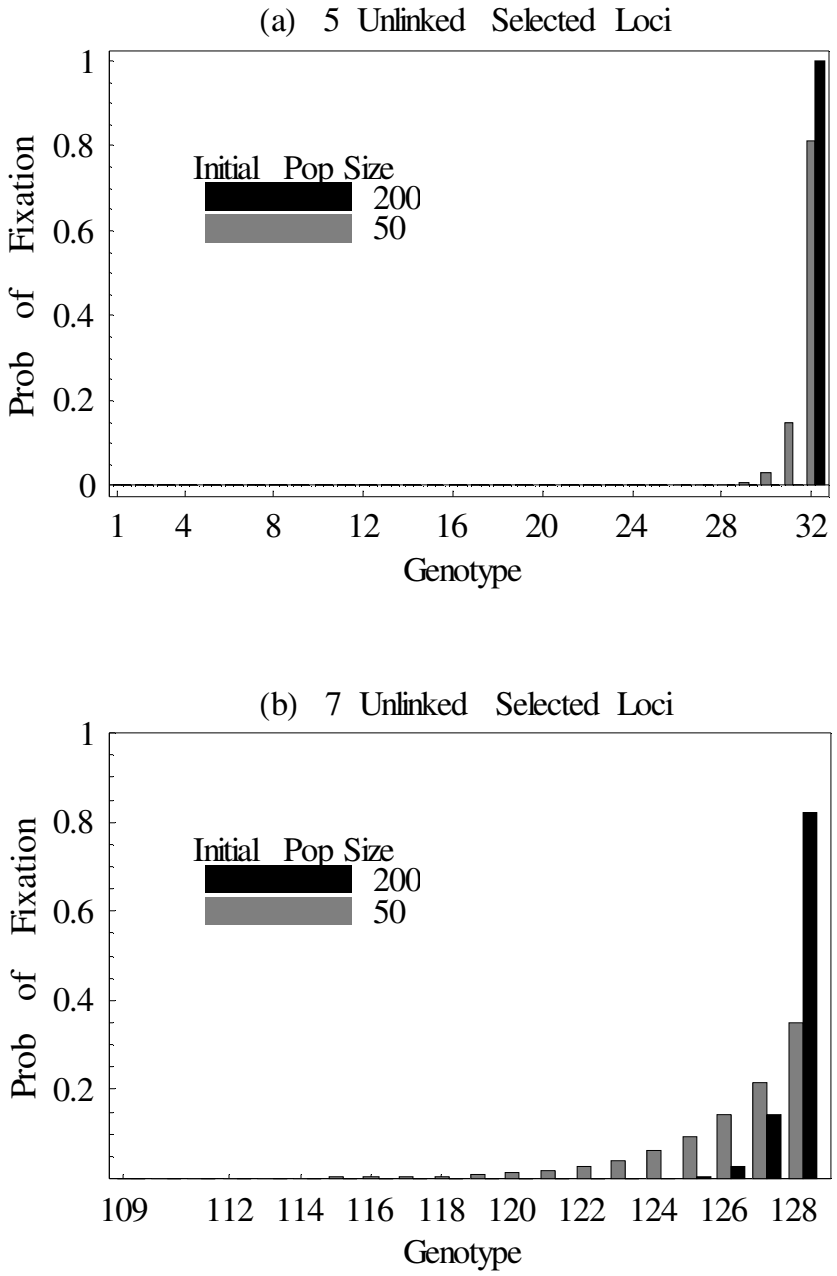


Figure 3.4 – Plot of the probability of fixation of the possible genotypes in a population. The gray bars represent the probability when the initial population size was 50, and the black bars represent the probabilities when the initial population size was 200. (a) is an example when there are 5 unlinked selected loci with effects {0.2, 0.05, 0.01, 0.03, 0.04}. (b) is an example where there are 7 unlinked selected loci with effects {0.2, 0.001, 0.1, 0.002, 0.15, 0.003, 0.02}.

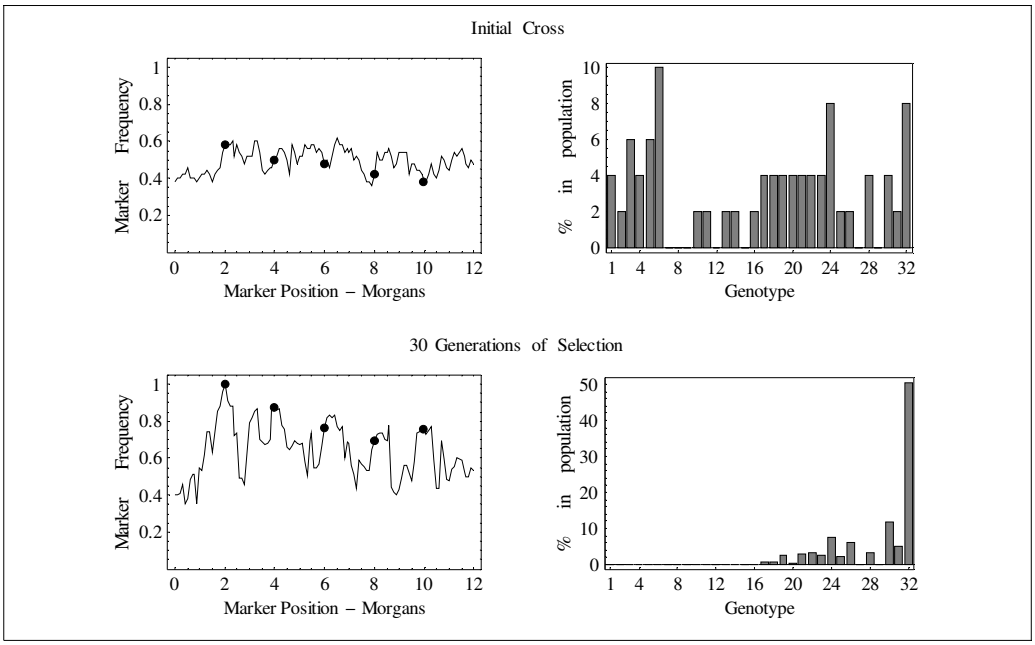
3.7 Variability in Genotype Fixation

Given that more than one genotype may have a non-zero probability of establishing in a population, what effect does this have on marker frequency? In any one replicate experiment the frequency of markers will look similar to that in Figure 3.2. That is, variation in frequency in unlinked regions will increase as selection is applied due to the reduction in the effective initial population size. However, the variance in frequency of selected alleles and associated linked markers may now be very different. When the same genotype established in each replicate, it was shown in Figure 3.3 that the variance in frequency of tightly linked markers and selected alleles reduced as selection was applied. However, with the possibility of different genotypes establishing, different alleles may fix at the selected loci in repeated experiments. Hence, the variance in frequency of selected alleles and tightly linked markers may now increase as selection is applied, mirroring the behaviour of unlinked markers.

Figure 3.5 shows an example of this. It shows two replicates of the marker frequency and genotypic composition of the population in an initial cross and after thirty generations of selection. The example is the same as the one in Figure 3.2, but the initial population size is now 50 instead of 200. Again, the positions of the five selected loci are indicated by the filled circles. With five unlinked selected loci and an initial population size of 50, the expected number of each genotype in the initial cross is just 1.56. With such a small expected number, in any one replicate, some genotypes will not be present in the initial cross. This is seen in Figure 3.5 where in both replicates there are missing genotypes in the initial cross. The fittest genotype, genotype number 32, is present in replicate 1, but is not present in replicate 2. As a result, after 30 generations of selection, there are different genotypes establishing in the two replicates. This results in different alleles fixing at the selected locus at position 6. It can be seen that the frequency at that selected locus is very high in replicate 1 but very small in replicate 2, similar to the pattern of some of the frequencies in unlinked regions. All the other selected loci have the same alleles fixing and hence have similar frequencies in both

replicates. So, overall it can be seen that if different genotypes can establish in different replicates, then the frequency at some of the selected loci may mirror the behaviour of frequencies in unlinked regions.

(a) Replicate 1



(b) Replicate 2

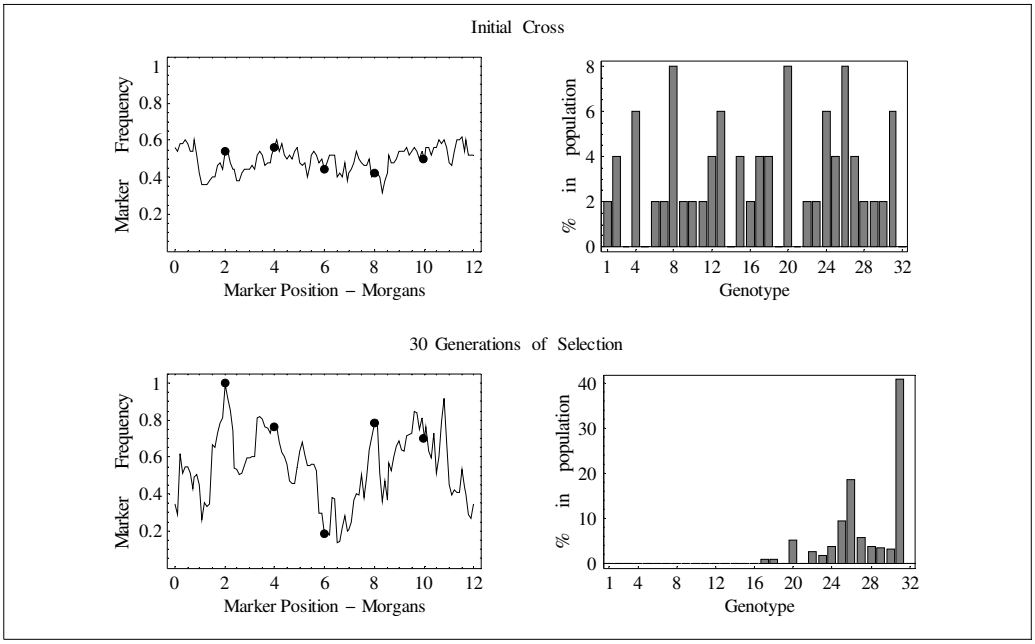


Figure 3.5 – Same as Figure 3.2, but the initial population size is now 50. With the lower initial population size, there is a much smaller probability that the fittest possible genotype will establish in each replicate. As a result variability in marker frequency may increase at some of the selected loci. This can be seen with the selected locus at position six.

3.8 Variation at Smaller Effect Loci

This increase in the variance in frequency at selected loci will generally only happen to the selected loci of smaller effect. This is because the genotypes that are fixing in each replicate are one of the genotypes in the upper tail of the fitness distribution. All these genotypes in the upper tail will have the fitter alleles from the large effect loci. The fitter alleles from the smaller effect loci, however, will be less abundant in these genotypes. Some of these genotypes will have the fitter alleles from the smaller effect loci, while others will not. Therefore, no matter what genotype establishes in the population, the alleles of large effect will always approach fixation, whereas with the smaller effect loci there is a greater probability that a less fit allele may fix.

Figure 3.6 shows a simple example of this. It plots the distribution of the fitter alleles across the 32 genotypes for the example in Figure 3.5. Figure 3.6(a) plots the distribution of the fitter allele from the large effect locus that was at position 2 in Figure 3.5, and Figure 3.6(b) plots the distribution of the fitter allele from the smaller effect locus that was at position 6 in Figure 3.5. It can be seen in Figure 3.6(a) that for the larger effect locus the fitter allele is always present in the genotypes in the upper tail. Hence, no matter which genotypes establishes in the population the fitter allele at locus 2 will always approach fixation. For the smaller effect locus at position 6, it can be seen that the fitter allele is distributed much more widely across the 32 possible genotypes, and as a result it is not present in some of the fitter genotypes. Consequently, as seen in Figure 3.5, if there is a possibility of different genotypes fixing in different replicates, then there will be variability in which allele fixes at these smaller effect loci.

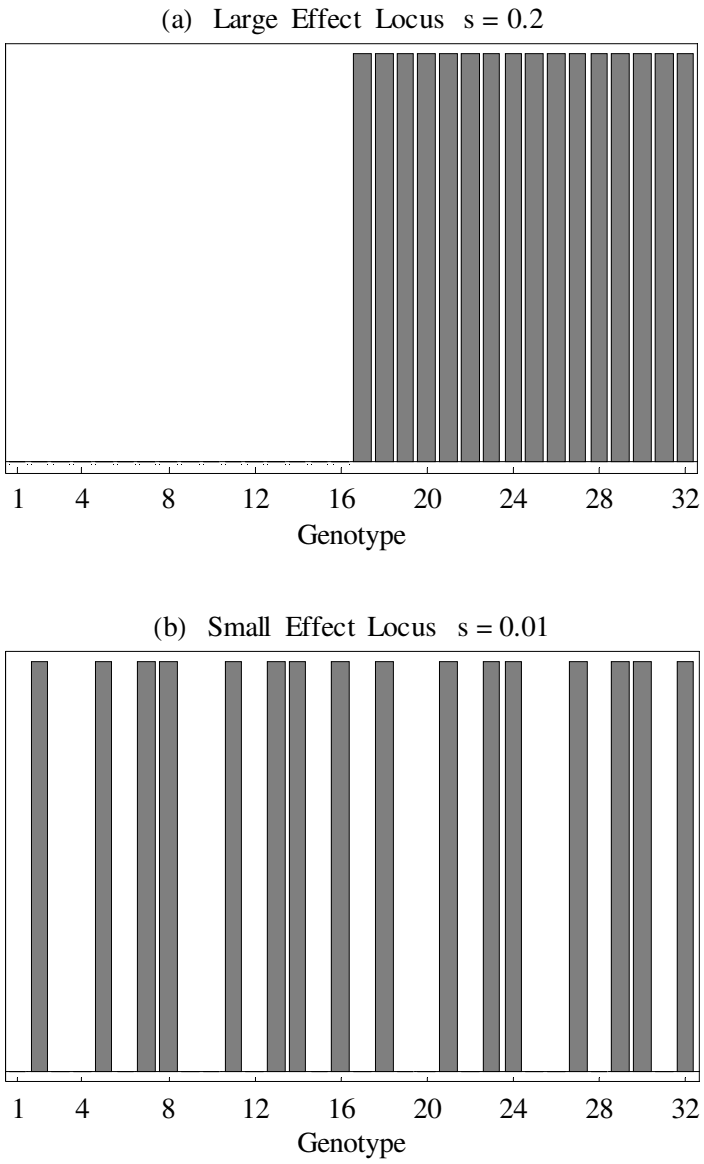
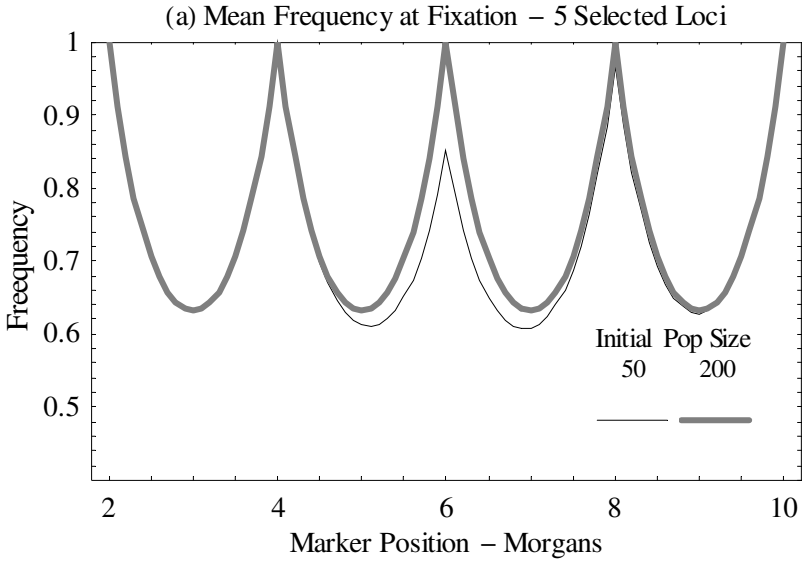


Figure 3.6 – (a) Plot of the distribution of the fitter allele across the 32 possible genotypes for the locus at position 2 in Figure 3.5. (b) Plot of the distribution of the fitter allele across the 32 possible genotypes for the locus at position 6 in Figure 3.5.

It is possible to get an idea of how much variability would be seen at these smaller effect loci due to different genotypes fixing, by looking at the mean frequency at fixation. So given that there are *n* fitness classes in the initial population, each having a probability

P_f^j of fixing, the mean frequency of a marker, if selection is continued to fixation is given by $\sum_{j=1}^{n} P_f^j E(F_m^j)$. Using this, Figure 3.7(a) plots the mean marker frequency at fixation for the example in Figure 3.5 when the initial population size is 50 and 200. It can be seen that with the smaller initial population of 50 there is one selected locus where there will be variability in which alleles fixes. This is the smallest effect locus. With a larger number of selected loci and/or a smaller initial population size, there would be many more selected loci which show variability. An example is shown in Figure 3.7(b), where there are seven selected loci with an initial population size of 200 and 50. It can be seen that there is a lot more variability at the smaller effect loci, and the average frequency at fixation of some of the smaller effect loci is near the null expectation of 0.5. So, overall we can see that if there is variability in which genotype establishes in each replicate, then the frequency of alleles at the smaller effect loci may start to resemble the pattern of behaviour of unlinked markers, but for the larger effect loci the same alleles will fix in each replicate.



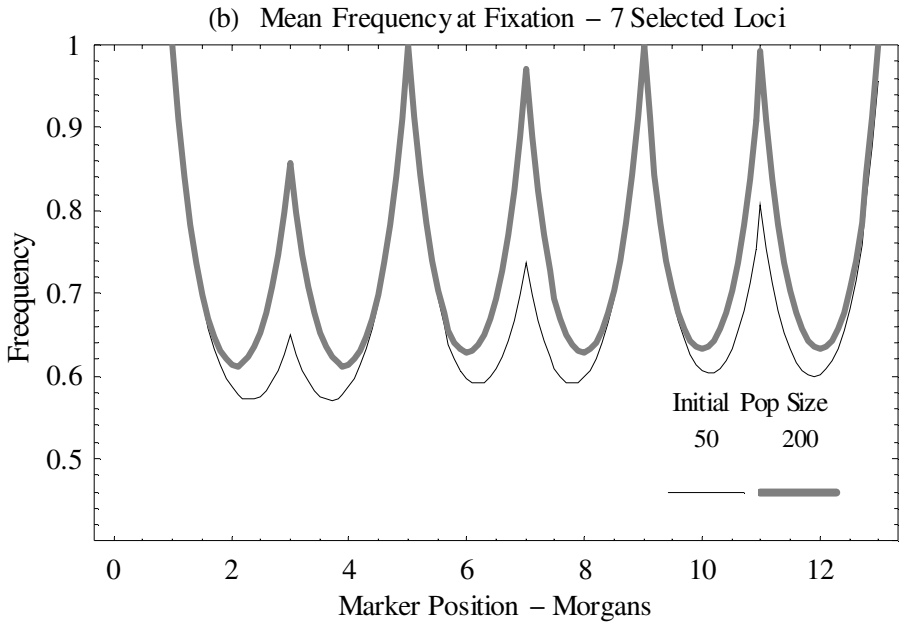


Figure 3.7 – Plot of the mean marker frequency at fixation, using $\sum_{j=1}^{n} P_f^j E(F_m^j)$, when the initial population size is 50 and 200, when there are five and seven unlinked selected loci. (a) The positions of the selected loci are {2, 4, 6, 8, 10}, and the selection coefficients were {0.2, 0.05, 0.01, 0.03, 0.04}. (b) The positions of the selected loci are {1, 3, 5, 7, 9, 11, 13}, and the selection coefficients were {0.2, 0.001, 0.1, 0.002, 0.15, 0.003, 0.02}.

3.9 Optimal Selection Time

In all the above examples which show the behaviour of the marker frequency, the most problematic case is the increase in the variance in frequency in the unlinked regions as selection is applied. The simplest solution to this is to increase the size of the initial population so that the variation in marker frequency would be reduced. However, as the number of selected loci gets larger the population sizes that are needed may become prohibitively large. Also, with a large number of selected loci, the fitness differences between the various genotypes may become quite small, and thus letting the experiment run until a genotype establishes in the population may not be feasible, as it may take a prohibitively long time for any one genotype to establish. So, overall it can be seen that

for quantitative traits, finding an optimal time to run the experiment in order get the maximum amount of information from the changes in marker frequency is necessary.

So how long should we select for? Ideally selection should be continued until such a time as the power to detect selected loci is maximized. One approach to do this could be to find a time when the frequencies of the selected alleles are high enough to be able to distinguish them from unlinked regions with high probability. Using this approach, the aim would be to find an optimal time t and a threshold, such that any marker frequency that has crossed the threshold can be assumed to have a high probability of being linked to a selected locus. The optimal time and threshold can be found by ensuring that markers that are closest to selected alleles will be more extreme in frequency than the extreme frequencies found in null regions.

Figure 3.8 shows an example of how this strategy may be implemented. It uses the example shown in Figure 3.2 where there were 5 unlinked selected loci all with different effects, and an initial population size of 200. I will assume that there are a total of 20 chromosomes, where each of the 5 selected loci are on separate chromosomes, and markers are equally spaced on each chromosome. Figure 3.8 plots, for each of the five selected loci, $P(u_{sel} > u_{thres})$, which is the probability that the frequency of the selected locus u_{sel}, is more extreme than a threshold frequency u_{thres} in unlinked regions, for each generation up to 200 generations. This threshold frequency u_{thres} is defined as $P(u_{\max} > u_{thres}) = 0.1$, where u_{max} is the maximum frequency in unlinked regions. For the example in Figure 3.8, both these probabilities were obtained from simulations. From the graph, it is possible to work out what the optimal selection time is for each selected locus. For example, for the largest effect locus, if we wanted to find a time such that there is at least a 90% probability that the frequency of the selected allele is more extreme in frequency than markers in unlinked regions, we see that this situation arises as early as generation 4. So, at this generation any marker frequency which is over the threshold u_{thres} for that generation can be assumed to be linked to that large effect locus.

This process can be repeated for each of the selected loci and a corresponding optimal time and threshold can be found for each locus. In order to obtain an overall optimal time and threshold to capture multiple selected loci, the selection time and threshold for the locus with the smallest effect is chosen. In Figure 3.8, if the cutoff was a 90% probability that the frequency of a selected locus is more extreme in frequency than markers in unlinked regions, then the overall time and threshold that would be chosen is the one associated with the locus with effect 0.04. The two smallest effect loci fail to reach the threshold after 200 generations of selection, and so if these two loci were also to be included then selection must be continued on for longer time and/or the size of the initial population must be increased.

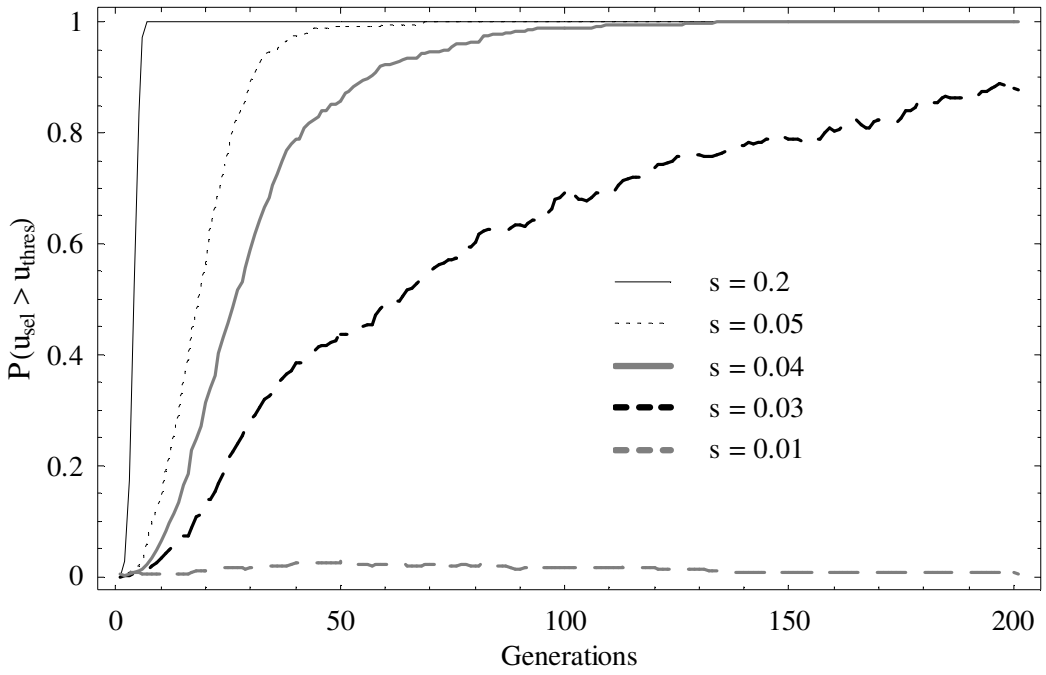


Figure 3.8 – For each of the five selected loci in the example in Figure 3.2, this graph plots, for each generation, the probability that the frequency of the selected allele is greater than the maximum frequency in unlinked regions. The probabilities were generated by simulations, where each selected locus was on a separate chromosome of length 1 Morgan, with a total of 20 chromosomes. There were ten equally spaced markers on each chromosome.

3.10 Setting a Time and Threshold

In the above example, very precise selection times and thresholds could be obtained as the number of selected loci involved and their effects were already known. Obviously, since the number of loci and their effects is the very information we are seeking to obtain from these experiments, this information would be not be available. So, in general it would not be possible to derive such precise times and thresholds as in the example above. So, in practice, if we were to use the same criteria as outlined in the example above to obtain selection times and thresholds, some assumptions must be made about the genetic architecture of the trait. Therefore using some assumptions, very general selection times and thresholds can be obtained.

For example, suppose we wanted to find selection times and thresholds such that there is a high probability of detecting loci of effect y and larger. To achieve this goal, a generation must be found such that a threshold frequency u_{thres} satisfies say, $P(u_{\max} > u_{thres}) = 0.1$ and $P(u_{sel} > u_{thres}) = 0.9$, where u_{max} is the maximum frequency in unlinked regions, and u_{sel} is the frequency of the selected allele. To evaluate these probabilities the distribution of marker frequency at generation t is needed. A normal approximation with moments provided by (2.6) – (2.8) and (3.1) – (3.3) can be used for this frequency distribution. However, to get the correct moments from these expressions, the number of selected loci and their effects are needed. Since this information is not available some approximations are needed. The simplest approximation to make is to evaluate $P(u_{\max} > u_{thres}) = 0.1$ and $P(u_{sel} > u_{thres}) = 0.9$ using a two fitness class model, where the difference in fitness between the two genotypes is the lowest effect locus we would like to detect. This model will give a reasonable approximation for the true threshold and selection times, provided that the initial population size used for the two fitness class model is equivalent to the effective initial population size of the correct

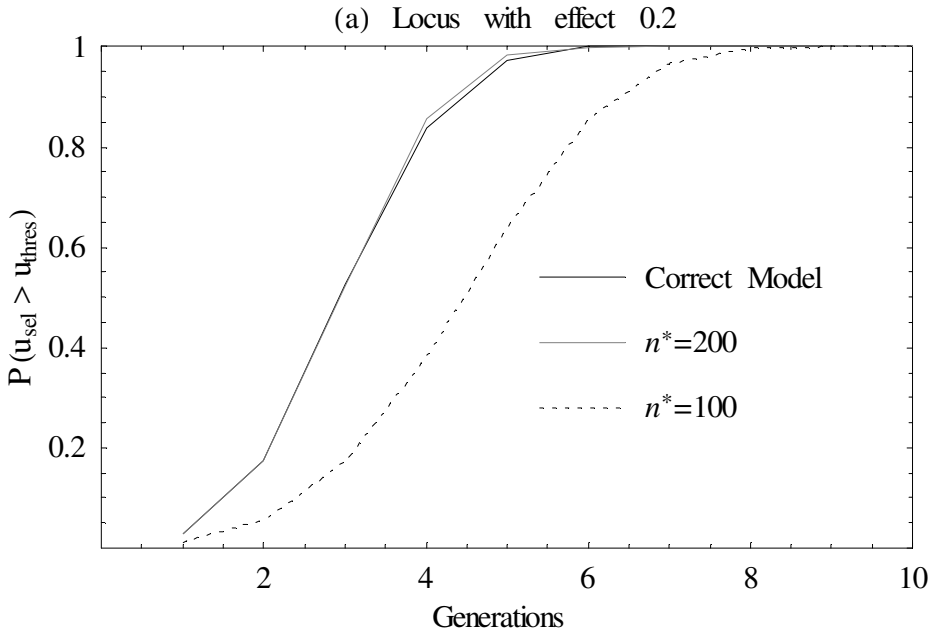
model. To work out the effective initial population size of the true model, we can assume that when the conditions $P(u_{\max} > u_{thres}) = 0.1$ and $P(u_{sel} > u_{thres}) = 0.9$ are satisfied, the locus with effect *y* is nearing fixation, and any higher effect loci are either fixed or also nearing fixation. Therefore, assuming there are *q* unlinked higher effect selected loci, the initial population size for the two fitness class model can be set to $n^* = 2^{-q} N$. Obviously, the *q* unlinked higher effect selected loci will again be an unknown, so some assumptions/guesses must be made here. So, using these parameters and the two fitness class model it is possible to evaluate thresholds and selection times. Specifically, a value for u_{thres} can be obtained by solving $P(u_{\max} > u_{thres}) = F_{CMVN}(\mathbf{u})^{c-1} = 0.1$, where $F_{CMVN}(\mathbf{u})$ is the cumulative multivariate normal distribution, **u** is a vector with all elements equal to u_{thres}, and of length equal to the number of markers on a chromosome, and *c* is the number of null chromosomes. Again *c* would be unknown so a rough idea of this is also needed, but the accuracy of this parameter *c* is not too important. Once the value for u_{thres} is obtained, $P(u_{sel} > u_{thres})$ can be evaluated. So, the generation and threshold in which say, $P(u_{sel} > u_{thres}) = 0.9$ is satisfied can be used as the optimal selection time and threshold.

Figure 3.9 shows an example of how accurate this approximation is. Figure 3.9(a) plots $P(u_{sel} > u_{thres})$ for the locus with effect 0.2 in the example in Figure 3.8. With an initial population of 200, and no loci with higher effect than 0.2, the parameters used for the two fitness class model were $y = 0.2$ and $n^* = (2^{-q})(N) = (2^{-0})(200) = 200$. The gray curve in Figure 9(a) is the two fitness class approximation result and the black curve is the result from the full correct model (ie. the same as in Figure 3.8). It can be seen that the two fitness class model provides a very good approximation. However, in practice it would not be known that there were no loci with higher effect than 0.2, so the dotted curve plots the result of the two fitness model if it was assumed that there is a single locus with higher effect. In this case the initial population size for the two fitness class model would be set to $n^* = (2^{-1})(200) = 100$. It can be seen that overestimating the

number of higher effect loci, overestimates the selection times. In this example the optimal selection time is only overestimated by a single generation. Figure 3.9(b) plots similar results for the locus with effect 0.05. Again, the gray curve is the two fitness class approximation result and the black is the result from the full correct model. There was one selected locus with higher effect, and hence the initial population size for the two fitness class model was set to $n^* = (2^{-1})(200) = 100$. Again, since the number of higher effect loci is an unknown, the other two curves in the graph show the result of the two fitness class approximation when the number of higher effect loci was underestimated and overestimated. The gray dotted curve plots the result when it was assumed there were two loci of higher effect, resulting in an initial population size of $n^* = (2^{-2})(200) = 50$ for the two fitness class model. The black curve plots the result when it was assumed there are zero loci of higher effect, resulting in an initial population size of $n^* = (2^{-0})(200) = 200$ for the two fitness class model. It can be seen Figure 3.9(b) that overestimating or underestimating the number of higher effect loci will produce very different selection times for this smaller effect locus. Underestimating the number of higher effect loci will result in a much earlier selection time, and consequently a higher false positive rate. Overestimating the number of higher effect loci will result in a longer selection time. In this particular example, the overestimation would not lead to an increase in the false positive rate, but it could do so if the effective initial population size was extremely low.

So, in general, it can be seen that obtaining accurate selection times and thresholds is very difficult unless precise information about the number of selected loci and the effects are known. The method outlined here would be most appropriate to use for large effect loci, as overestimating or underestimating the number of higher effect loci will have relatively little adverse effect on the selection times and false negative rates. For smaller effect loci, however, it can be seen from Figure 3.9(b) that the predicted selection times can be very different from the optimal times, and so unless accurate

estimations are available about the number of higher effect selected loci, then this method may give very misleading results.



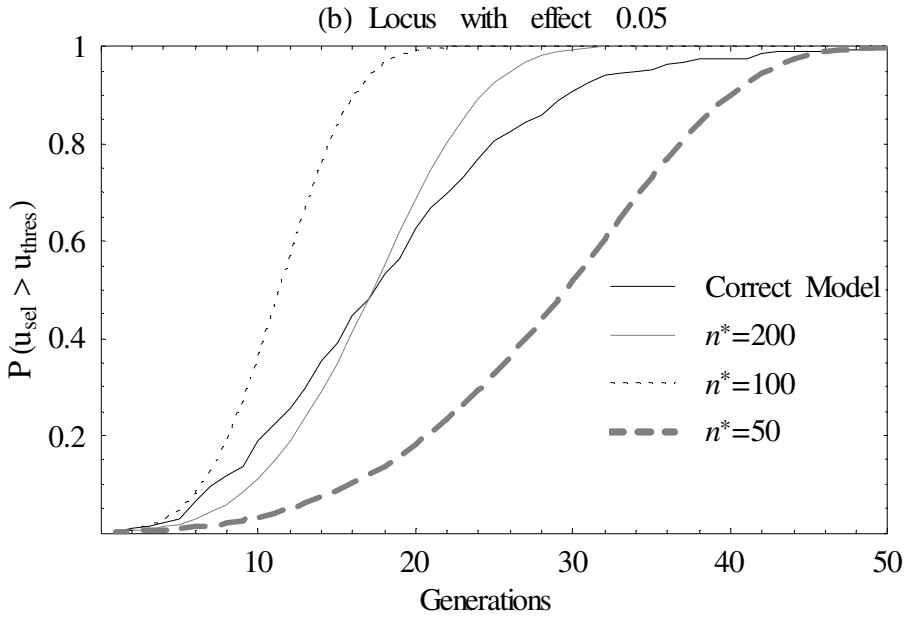


Figure 3.9 – This figure plots the probability that the frequency of the selected allele is greater than the maximum frequency in unlinked regions using a two fitness class approximation for two of the selected loci shown in the example in Figure 3.8.

3.11 Epistasis

So far an additive model has been analysed where the effect each allele has on fitness was independent of other selected loci. This meant that an allele had the same effect on fitness no matter what genetic background it was on. If, however, there was interaction between selected loci, and an allele's fitness depends on its genetic background, then at any particular generation the marker frequencies at the selected loci may be very different from an additive model. If selection is continued for long enough, we would still expect to see peaks and valleys in the marker frequency around the selected loci in the selected population, but which alleles fix at these selected loci and at what times these alleles fix may be very random. Consequently, it may be difficult to identify what effect each of the selected alleles has on the selected trait. This behavior will first be outlined with a two locus model and then outlined when interaction occurs between many selected loci.

3.12 Epistasis – Two Loci

Suppose there are only two selected loci, locus A and locus B, and assume that both loci have two possible alleles, a_1, a_2 and b_1, b_2. Suppose alleles a_1 and b_1 are the fitter alleles and they have an effect s_1 and s_2 on fitness where $s_1 > s_2$. The four possible genotypes and their fitnesses are shown in Table 3.1. The last column in Table 3.1 shows the fitnesses when it is assumed that there is interaction between alleles a_1 and b_2. This results in the fitness of genotype a_1b_2 changing from $(1+s_1)$ to $(1+s_1)+\varepsilon$, where ε represents the epistatic effect.

Genotype	Fitness – Additive	Fitness – Epistasis
a_1b_1	$(1+s_1)(1+s_2)$	$(1+s_1)(1+s_2)$
a_1b_2	$1+s_1$	$(1+s_1)+\varepsilon$
a_2b_1	$1+s_2$	$1+s_2$
a_2b_2	1	1

Table 3.1 – This table shows the four possible genotypes and their fitness, in and additive and epistasis model, when there are two selected loci. It is assumed $s_1 > s_2$.

How will this affect marker frequency? Figure 3.10 shows the pattern of marker frequency that would be expected after ten generations of selection for different values of ε. In the example, locus A is at position 3 with $s_1 = 0.15$ and locus B is at position 7 with $s_2 = 0.05$. Firstly, under an additive model (ie. $\varepsilon = 0$), it can be seen that, as expected, the frequency of markers around the selected alleles increase, with a larger change in frequency with the larger effect allele. With epistasis involved, ε will now take on a positive or negative value. When $\varepsilon < 0$ the fitness of genotype a_1b_2 is reduced, and so the overall effect of alleles a_1 and b_2 on phenotype will be reduced in the population. This effectively results in altering the fitness differences between the alleles at the selected loci. That is, when $\varepsilon < 0$, in effect, the fitness difference between the alleles at locus A is reduced and the fitness difference between the alleles at locus B is increased. Consequently the time to fixation/loss of the alleles at the selected loci will be different from the additive model. This is shown in Figure 3.10 when $\varepsilon = -0.09$, where the frequency of allele a_1 is slightly lower, and the frequency allele b_1 is much higher.

The next case is when ε is positive. When $\varepsilon > 0$ the fitness of genotype a_1b_2 increases. In this case there are two possible outcomes. Firstly the case when $0 < \varepsilon < s_2(1+s_1)$. This is the situation when the fitness of genotype a_1b_2 increases but still less then the fitness of the fittest genotype in the additive model a_1b_1. Once again, epistasis results in a

change in the fitness differences between the alleles at the selected loci. In this case the fitness differences between the alleles at locus A will increase and the fitness differences will decrease for alleles at locus B. In Figure 3.10 $\varepsilon = 0.05$ shows this possibility. The increase in the fitness difference at locus A results in a slight increase in the frequency of allele a_1. For allele b_1 the frequency is now greatly reduced as the interaction has caused the fitness differences between the alleles at locus B to become very small. Hence selection must be applied for much longer to see an appreciable change in frequency in the alleles at this locus B. The other case when $\varepsilon > 0$ is when $\varepsilon > s_2(1+s_1)$. In this case the fittest genotype is no longer a_1b_1, as the genotype a_1b_2 will now have a larger fitness. So, now at locus B, allele b_2 will establish in the population instead of allele b_1. In terms of time to fixation of these selected alleles a_1 and b_2, the pattern is similar to the previous cases. The frequency of allele a_1 will be higher than the additive model as the fitness differences at that locus have increased. For locus B, the allele b_2 will eventually fix, but the time to fixation will depend on the exact value of ε. Figure 3.10 shows the case when ε is quite large at $\varepsilon = 0.35$. In this case we see that allele a_1 is nearing fixation and fitter allele b_1 is approaching a frequency of zero.

So, overall it can be seen that interaction between selected alleles alters the fitness of genotypes, and consequently the frequencies of the selected alleles may be quite different from an additive model. From Figure 3.10, the most problematic case when interaction alters fitness is when the fitness difference between the alleles at a selected locus become very small. In this case, the effect that a locus has on phenotype may be masked, unless selection is continued on for a very long time.

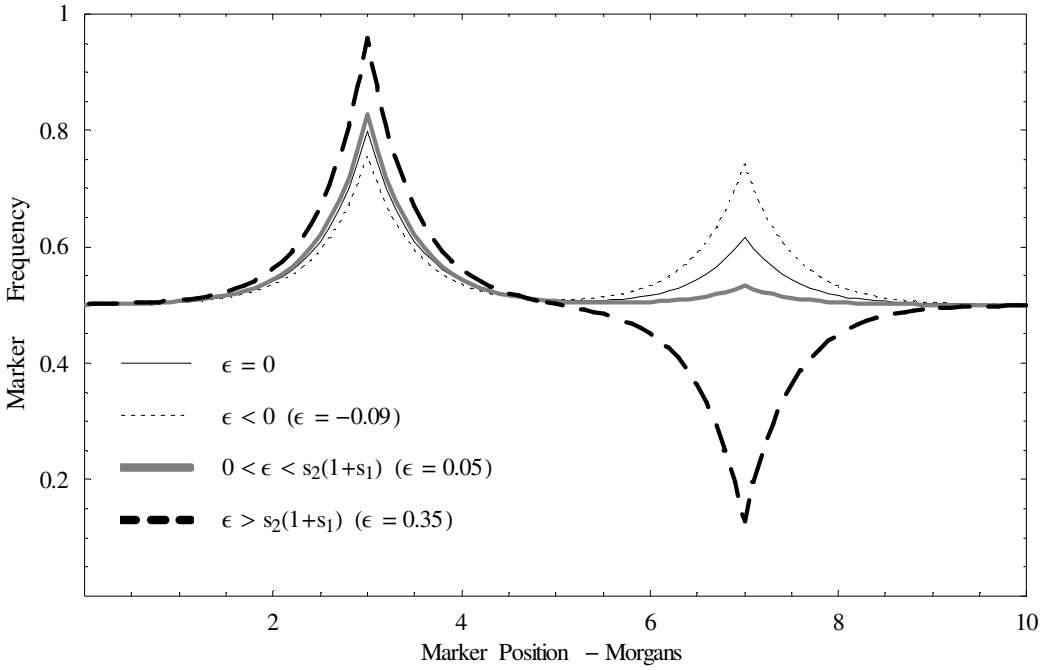


Figure 3.10 – Plot of the expected marker frequencies (using (3.1) – (3.3) and (2.6)) after ten generations of selection under an additive model and epistasis model when there are just two selected loci. The selected loci are at position 3 and 7, with the selected alleles having effects 0.15 and 0.05. The additive model is the graph $\varepsilon = 0$, whereas the other three graphs show the different possibilities when interaction alters the fitness of a genotype.

3.13 Epistasis – Multiple Loci

With more than two selected loci there are a lot more possibilities for interaction between the selected alleles. Consequently, there could be many alleles that differ in frequency from an additive model. In fact, the larger the number of interactions, the greater the probability that the fitness differences between alleles at a selected locus will become very small. Consequently, the frequency of alleles at selected loci will most likely reflect the situation in Figure 3.10 at locus B (position 7) when $\varepsilon = 0.05$. This is the situation where there is very little change in frequency at the selected locus. This is because, with a large number of interactions, the fitness of genotypes may have no correlation with the alleles present at the selected loci. So, the fittest genotype in the

population may have no resemblance to the next fittest genotype and so on. Hence, with no consistency in which alleles confer a fitness advantage, the result will be a much slower change in frequency at selected loci. This makes identification of selected loci increasingly difficult.

This can be illustrated with the extreme case when interaction causes the fitness of all recombinant genotypes to be different from an additive model. To illustrate this, I will again use the five selected loci example used throughout this chapter (example detailed in Figure 3.1). Figure 3.11(a) plots the expected marker frequency (using (3.1) – (3.3) and (2.6)) after thirty generations of selection in the additive case. For the epistasis case, to simulate the large number of possible interactions between the five selected loci, each of the 32 possible genotypes was assigned a relative fitness from a uniform distribution ranging from 1 to 1.36 (the range was chosen as 1 to 1.36 because this was the range in the additive case). Figures 3.11(b) and 3.11(c) plot two examples of the expected marker frequency after thirty generations of selection in the epistasis case. It can be seen from the two epistasis examples, the frequencies at the selected loci are not as extreme as in the additive case. In the additive case there was one very large effect locus at position 1 which results in the expected frequency of the selected allele at that locus to be 0.99 at generation 30. From the two epistasis examples shown in Figure 11, the most extreme expected frequency of any of the selected alleles is the selected allele at position 10 in replicate 1, whose frequency is 0.28, which is only a 0.22 change in frequency from the null expectation of 0.5. This slower change in frequency is due to the smaller fitness differences at the selected loci, as a result of the random distribution of the selected alleles on the fitter genotypes. Figure 3.11(d) plots the distribution of the expected change in frequency from the null expectation of 0.5, for the selected allele at the most extreme frequency in a 1000 replicates in the epistasis model. From Figure 3.11(d) the average frequency of the most extreme selected allele after thirty generations of selection is just 0.23 from the null expectation of 0.5. So, it can be seen that if there is a very large amount of interaction between the selected alleles, then selection may need to be applied for a much longer time in order to see any appreciable changes in marker

frequency. However, as previously discussed, the longer selection is applied the larger the stochasticity in frequency in unlinked regions. Hence, with large numbers of interacting loci, unless very large initial population sizes are available and selection is continued for a very long time, it may be very difficult identify any selected locus from marker frequency changes.

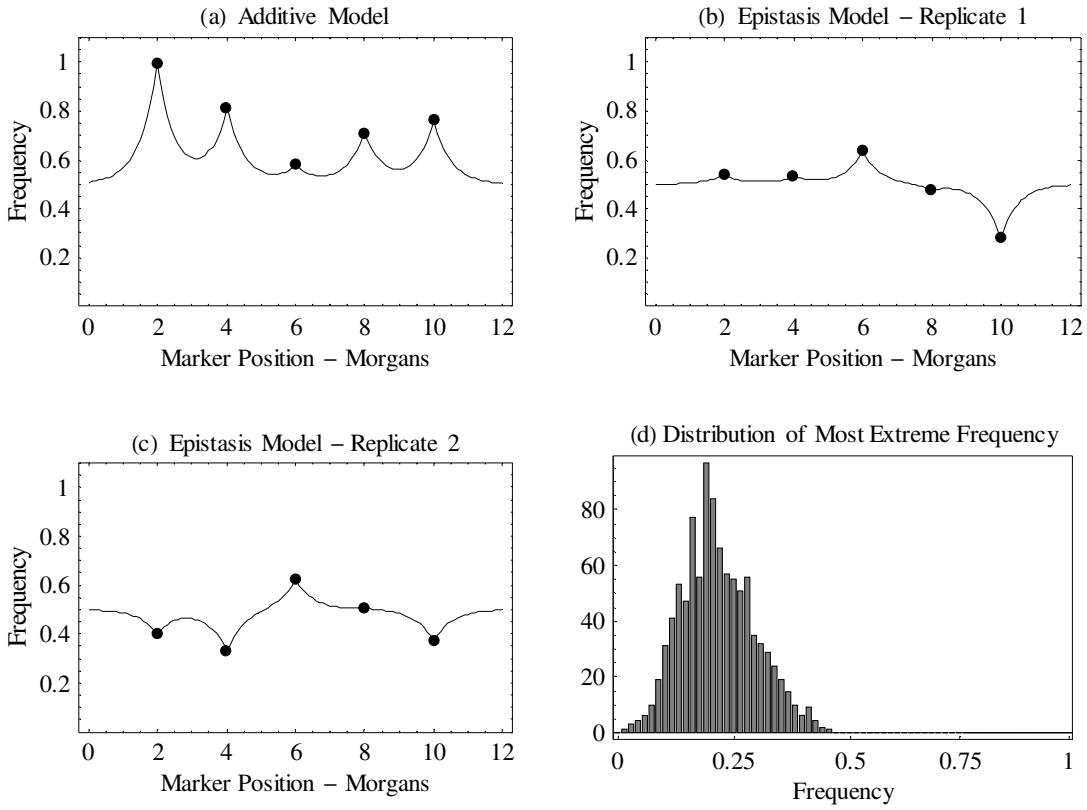


Figure 3.11 – (a) plots the expected marker frequencies (using (3.1) – (3.3) and (2.6)) after thirty generations of selection when there are five selected loci. The positions of the selected loci are shown by the filled circles. The five selected loci have effects {0.2, 0.05, 0.01, 0.03, 0.04}. (b) and (c) show the expected marker frequencies after thirty generations of selection when there is interaction between the five selected loci. Each genotype was assigned a fitness from a uniform distribution. (d) plots the distribution of the expected change in frequency from the null expectation of 0.5, for the selected allele at the most extreme frequency over a 1000 replicates in the epistasis model

3.14 Detecting Epistasis

If marker frequencies are obtained from just a single generation, then it would not be possible to tell from just analyzing the frequencies whether there is interaction occurring between selected loci. However, in some cases, if marker frequencies from several generations are available, then it may be possible to tell if interaction is occurring. This is because in some cases epistasis may result in the frequency change of selected alleles reversing in direction. So, if some markers are initially observed decreasing in frequency, but in later generations increasing in frequency (or vice versa) then this would indicate interaction between selected loci. This is due to the way selected alleles are distributed on the fitter genotypes in the population. For example, in an additive model, many of the genotypes in the upper tail of the fitness distribution will have same alleles at the selected loci. This can be seen in Figure 3.6. This results in the frequency of alleles at the selected loci always either increasing or decreasing once selection is applied. However, in a model with epistasis, there may be situations where the fittest genotype in the population has alleles at selected loci that the vast majority of the other fitter genotypes do not have. In this case the direction of the frequency change of alleles at these selected loci will change in the later generations. An example of this is illustrated in Figure 3.12 using the epistasis setup used in Figure 3.11(b) and Figure 3.11(c). Figure 3.12(a) plots a possible distribution for a selected allele across the 32 possible genotypes under an epistasis model. It shows an extreme example where a particular allele at a selected locus is present on the fittest genotype (genotype 32), but this allele is missing on all the other genotypes in the upper half of the fitness distribution. Using (3.1) – (3.3) and (2.6), Figure 3.12(b) plots the mean frequency of this selected allele over time. It can be seen that the frequency initially decreases, but then increases as the fittest genotype increases in numbers in the population. This is confirmed by simulations shown in Figure 3.12(c) – 3.12(f), where the selected locus in question is at position 10. It can be seen that after 20 generations of selection there is a single clear drop in marker frequency around position 10 which would indicate the presence of a selected locus. However, after 50 generations of selection the drop in

frequency at position 10 disappears and the frequency at the locus is back around the null expectation of 0.5. After 100 generations of selection the frequency change at position 10 is clearly in a different direction to the earlier generations as it is now approaching fixation. So, this reversal pattern of marker frequency can be used as an indication of the presence of interaction between selected loci. However, the absence of such a pattern would not indicate an additive model, as interaction between selected loci can easily still occur without such a pattern.

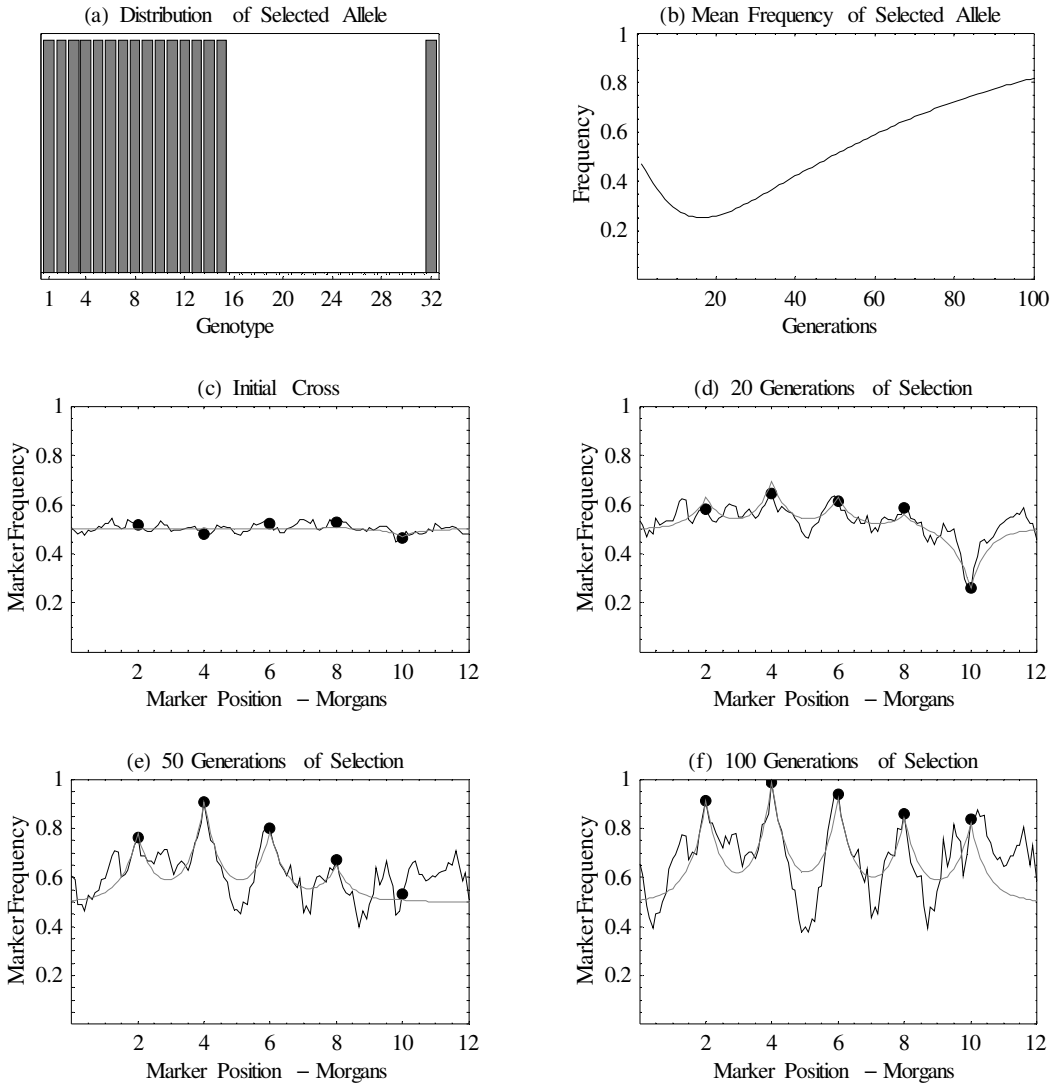


Figure 3.12 – This figure illustrates an example where the direction of frequency change of selected alleles reverse as selection is applied in an epistasis model. The model for the epistasis was the same as used in Figure 3.11(b) and Figure 3.11(c). That is, five unlinked selected loci, where each of the 32 possible genotypes were assigned fitnesses from a uniform distribution ranging from 1 to 1.36. (a) Shows an example of a possible distribution for a selected allele in such a model. (b) Plot of the mean frequency against time for the selected allele shown in (a). (c) – (f) Plot of simulation results of the marker frequency at various stages of selection. The selected allele depicted in (a) and (b) is at position 10. The light gray curve in (c) – (f) is the deterministic expectation for the marker frequency.

3.15 Discussion

In this chapter the aim was to identify alleles that influence a quantitative trait by examining changes in marker allele frequency in selected progeny. Once again, the extreme progeny are selected by multiple generations of asexual reproduction and selection. With quantitative traits, where multiple loci influence trait value, the initial population will consist of a distribution of fitness classes. This is in contrast to the previous Mendelian trait model, where only two fitness classes were present in the initial population, and it was assumed that only the fitter class of recombinants survived in the selected population. However, it was shown in this chapter that selecting out only the fittest class of recombinants (or any single fitness class) would most likely not be desirable, as it can lead to large stochasticity in marker frequency in the selected population. This is because when many loci influence the trait, each fitness class would most likely be at low numbers in the initial population due to low probability of being produced at meiosis. As a result, if the selected population consisted of only descendents from a single fitness class, then there would be large stochasticity in marker frequency in unlinked regions. Hence, selecting until a genotype fixes would in most cases not be appropriate, and finding optimal selection times would be more desirable. However, it was shown that finding an optimal selection time is a difficult process. For any particular criteria for optimal selection times, correct information about the number of selected loci and their effects need to be known in order to get accurate selection times. Since this information is obviously not available, some estimations about the number of selected

loci is needed, and as a result only very general selection times can be obtained. For the criteria that was used in this chapter to find optimal selection times, it was shown that overestimating or underestimating the number of selected loci can give very misleading selection times which could lead to many false positives. Therefore, if selection is going to be applied for a very long time, then extremely large initial population sizes should be present in order to minimize the variance in unlinked regions, and as a result reduce the likelihood of false positives.

The other issue with quantitative traits is the role of epistasis. Assuming a large enough initial population size is present, and selection is continued for long enough, then it was shown that epistasis will generally pose no problems. That is, clear peaks and valleys in the marker frequency around the selected loci will appear in the selected population similar to any additive model. However, if there is interaction between large numbers of selected loci, then one problem that may exist is the length of time it may take to get an appreciable change in frequency at the selected loci. It was shown that selection may need to be continued on for a much longer time to see changes in frequency at the selected loci. So, in this case, large initial population sizes would be a necessity, as selection would have to be applied for a very long time to get any signal. The final issue that was analysed in this chapter was how to detect the presence of epistasis from the frequency of markers. It was shown that for some models of epistasis the direction of frequency change of markers linked to some selected loci will reverse as selection is applied. So, if marker frequencies from several timeframes are available, then identifying such a pattern would indicate interaction between selected loci.

Chapter 4: Data Analysis

Abstract

Most linkage analysis experiments involving experimental crosses rely on marker phenotype correlations to infer linkage between markers and quantitative trait loci (QTL). As a result, the majority of statistical methods developed to detect QTL from linkage analysis experiments are based on analysing phenotype data. These statistical tests can be generally divided into methods that search for a single QTL at a time and methods that simultaneously search for multiple QTL. In this chapter, I will concentrate on statistical methods that search for a single QTL at a time. Firstly, the main single QTL mapping methods that analyse phenotype data will be outlined. I also give a very brief review of the few statistical methods that exist to analyse marker frequencies in pooled selected asexual cross progeny. I will then outline a method, similar to standard interval mapping, which searches for a single QTL at a time in pooled selected asexual cross progeny based on the branching process model. It will be shown, using simulated data, that when a selected allele has fixed in the population, this model will successfully identify the general location of that QTL given that the effective initial population size N^* is not extremely small. The accuracy of the location estimate increases with the size of N^*. Finally, the statistical model is applied to some data obtained from a malaria experiment.

4.1 Introduction

4.1.1 Comparing Phenotype Means

When analyzing marker phenotype correlations, and searching for a single QTL at a time, the simplest method is to analyse each marker locus separately, and test whether the locus is linked to a QTL. In order to do this, for each marker locus, the mapping population is divided into groups based on alleles at the locus. To test for linkage, the trait means of the various groups at a particular locus are compared. If there is a significant difference in the trait means at a marker locus, then that marker locus can be assumed to be linked to a QTL. Significant differences between trait means are usually determined by a *t*-test. This process is continued for each marker locus. This is a very simple method which can identify linkage between marker and a QTL, but its main drawback is that it can be difficult to get a precise location of the QTL. This is because it is not possible to separate to the effect of the QTL from its position. That is, a significant difference between trait means may indicates linkage, but it is not possible to tell whether this linkage it due to a nearby QTL of small effect or a far away QTL of large effect.

4.1.2 Interval Mapping

The most common way to overcome the problems associated with just comparing phenotypic means is to specifically model the location of the QTL with respect to the markers, enabling separate estimates of QTL location and effect. This is generally done by the method of maximum likelihood, where the distribution of trait values are compared in a model with a QTL present, to a model where there is no QTL present. This methodology can be applied to a single marker at a time, but it is usually applied using two markers at a time and is known as interval mapping (Lander & Botstein, 1989).

With interval mapping, two markers at a time are analysed on each chromosome. For each pair of markers that are analysed, a likelihood ratio is calculated. The hypotheses that are tested by the likelihood ratio are, that a single QTL exists somewhere between the two markers being analysed, and the null hypothesis that no QTL exists between the two markers. For the null hypothesis it is assumed that the phenotype values y are Gaussian distributed, $f(y;\mu,\sigma^2)$ with mean μ and variance σ^2. Therefore the likelihood function for the null hypothesis is simply $\prod_i^n f(y_i;\mu,\sigma^2)$, where n is the number of individuals, and the unknown parameters that need to be estimated from the data are μ and σ^2. For the alternate hypothesis, it is again assumed that the phenotype values are Gaussian distributed, $f(y;\mu_j,\sigma^2)$ where μ_j is the mean effect of QTL genotype j between the two markers being analysed ($j=2$ for a backcross population and $j=3$ for an F_2 population). Since the QTL genotype between the two markers is generally unknown, the phenotype distribution is set as a normal mixture distribution using the genotypes at the two flanking markers. So, for any two particular markers, the phenotype distribution can be written as $\sum_j p_{ij} f(y_i;\mu_j,\sigma^2)$, where p_{ij} is the probability that QTL genotype j exists between the two markers given the marker genotype data of individual i. Consequently, the likelihood function for the alternate hypothesis is $\prod_i^n \sum_j p_{ij} f(y_i;\mu_j,\sigma^2)$, where the unknown parameters that need to be estimated from the data are the μ_j, σ^2, and the location of the QTL between the two markers (the location parameter would be embedded in p_{ij}).

So, using these distributions and estimates of the unknown parameters, a likelihood ratio is calculated for each pair of markers along the genome. Once this is completed, a linkage map score is constructed, which is a plot of the likelihood ratios against map position. Any significant peaks in the linkage map are indications of the presence of QTL. The threshold for the significance in the linkage map is usually determined by asymptotic results or permutation analysis. Both methods essentially find a threshold by

deriving a distribution for the maximum value of the genome wide likelihood ratio under the null hypothesis, and extract a threshold from this distribution.

4.1.3 Regression Mapping

The advantages of this interval mapping technique are that separate estimates of QTL position and effect can be obtained. However, one of the main drawbacks is the increase in computation time in implementing the method. Estimating the unknown parameters in the likelihood functions is usually done by an iterative algorithm, such as the expectation-maxmisation algorithm, which can be time consuming. One way to overcome these issues is to approximate the interval mapping technique by using regression mapping, which gives similar mapping power, but reduced computation time (Haley & Knott, 1992).

With regression mapping, it is assumed that for each pair of markers the phenotype values y are Gaussian distributed, $f(y; p_{ij}\mu_j, \sigma^2)$ with mean $p_{ij}\mu_j$, and variance σ^2, where μ_j and p_{ij} are defined the same as in the interval mapping case. Since the expectation of this distribution is a linear function of μ_j, the μ_j can be estimated by regressing the phenotype y on the probabilities and p_{ij}. Therefore, for each interval that is tested, computation times are reduced as the unknown parameters are estimated by a simple linear regression.

4.1.4 Marker Frequency Data – Asexual Cross Progeny

When the data consist of marker frequencies obtained from selected pools of asexual cross progeny, the statistical methods that have been developed to identify QTL have mainly been based on using *t*-tests. These tests involve obtaining several replicates of

selected and unselected pools and comparing the mean frequency of markers in the two pools.

Specifically, in the method outlined in (Segre *et al.*, 2006) they define a linkage likelihood ratio $LLR = \int_{t_1}^{\infty} g(t)dt \Big/ \int_{t_2}^{\infty} g(t)dt$ for each marker in the genome. In this ratio, $g(t)$ is the t probability distribution, t_1 is the t test statistic obtained when comparing the mean intensity of a marker in the selected pools to the mean intensity of the marker in the parental strain, and t_2 is the t test statistic obtained when comparing the mean intensity of a marker in the selected pools to the mean intensity of the marker in the unselected pools. Once this ratio is calculated for each marker on the genome, a sliding window is applied along the genome, where within each window the geometric mean of the *LLR* of the markers within the window are calculated. Any geometric mean over a certain threshold is assumed to be a QTL, where the threshold is obtained by permutation analysis.

The other method is the method outlined in (Ehrenreich *et al.*, 2010). In this method t tests are again used, but QTL are now identified by searching for inflection points in the p-values associated with each marker. Specifically, for each marker the mean frequency in the selected pools is compared to the mean frequency in the unselected pools using a t test, and the resultant p-value is recorded. A sliding window is then moved along the genome and the average of the $-log_{10}(p)$ values within each window is recorded. Using these averaged p-values, a further sliding window is moved along the genome, where within each window linear regressions are fitted and the slope of the regressions are recorded. An inflection point is defined as a point where the slope changes sign. So, within any window, whenever the slope of the regression changes sign, that position is assumed to be the position of a QTL, provided that the average $-log_{10}(p)$ value associated with that window is over a certain threshold. Again thresholds are obtained by permutation analysis.

4.2 Branching Process Model

The advantages of the *t*-test methods are that they are relatively easy to apply, make no assumptions about the number of QTL involved, and they have shown they can accurately map many QTL. However, one disadvantage of these methods is that several replicates of unselected and selected populations are needed to implement the *t*-tests. One cannot assess significance in marker frequency by applying those methods to a single selection experiment. In this chapter, a maximum likelihood method based on the branching process will be outlined to assess the significance in marker frequency in selected pools of asexual cross progeny. Since, the branching process enables us to derive specific genetic models for the distribution of marker frequency in the selected population, the significance of marker frequencies can be assessed within a single selection experiment. I will outline this branching process statistical method for the simplest genetic model of a single fully selected locus. The aim is to get a statistical estimation of the location of selected alleles that have fixed in the population.

4.2.1 Single Fully Selected Locus

The methodology is similar to many QTL mapping methods in that the genome is sequentially scanned for selected loci. The genome is divided into intervals of c cM (where the intervals could be overlapping), where c is typically 10 – 20, and each interval is tested for the presence of a selected allele. For each interval that is tested, the two markers that define the interval are used for the likelihood analysis. This results in the likelihood setup being similar to the standard interval mapping setup. More markers could be used for the likelihood analysis for each interval, but it will be shown later that using more markers does not improve the mapping accuracy. So, in general, the mapping process is that for each interval, the two markers that define the interval are used to calculate a log likelihood ratio $\lambda = Log(L_0 / L_A) = Log(L_0) - Log(L_A)$. L_A is the

likelihood under the hypotheses that a single selected allele is fixed somewhere on the interval being tested, and L_0 is likelihood under the null hypothesis that no selected allele exists on the interval being tested.

4.2.2 Likelihood Functions – Gaussian Approximation

The likelihood functions for both L_0 and L_A can be obtained by using the distribution of marker frequency from a genetic model of a single fitness remaining in the selected population, where all individuals carry the selected allele. It was shown in Chapter 2 that this distribution, given that the size of the initial population is not too small, can be approximated by a k-variate Gaussian $f_k(m,\Sigma;y) = (2\pi)^{-k/2}\det(\Sigma)^{-1/2}\exp(0.5(y-m)'\Sigma^{-1}(y-m))$, where k refers to the number of markers being used in the analysis, m is a vector of the theoretical mean frequencies of the k markers, Σ is the theoretical covariance matrix, and y is the data vector containing the marker frequencies from the experiment. The mean vector $m_i = P_i$, where P_i is the probability that marker i is on an individual in the initial population, where the initial population consists of an expected number of n individuals, all of which carry the selected allele. $\Sigma_{ij} = V * C_{ij}$, where $V = 1/n(1+(\sigma^2(1-\mu^{-t})/\mu(\mu-1)))$ and $C_{ij} = P_{ij} - P_i P_j$, where P_{ij} is the probability that marker i and marker j are on an individual in the initial population. The log likelihood of f_k for use in the likelihood ratio is given by (4.1).

$$Log(f_k) = l(x,V) = -0.5((y-m)'\Sigma^{-1}(y-m) + Log(\det(\Sigma))) \qquad \textbf{(4.1)}$$

To calculate the likelihood ratios from (4.1) values for the unknown parameters need to be obtained. The two unknown parameters in (4.1) are the constant V and the location of the selected locus x. The location parameter x would be embedded in the function used

for the recombination probabilities for the elements in *m* and *C*. Outlined below is the procedures for estimating both these parameters.

4.2.3 Maximum Likelihood Estimator for V

A maximum likelihood estimator for *V* can be obtained by setting $\Sigma = V * C$ in (4.1), and solving $dLog(f_k)/dV = 0$ for *V*. Doing so, we get $\hat{V} = ((y-m)'C^{-1}(y-m))/k$ as the maximum likelihood estimator for *V*. There are two ways in which to use this estimator. For each interval that is tested, a separate estimate of *V* can be obtained, or a single global estimate of *V* can be obtained, and this global estimate can be used for each interval that is tested. If separate estimates of *V* are to be obtained for each interval, it would mean obtaining separate significance levels for each interval. However, a global estimate of *V* would mean using a genome wide significance level. As interval specific significance levels typically result in a higher false positive rate across the genome, we will seek to obtain a global estimate of *V*, and consequently also a genome wide significance level.

4.2.4 Estimating a Global V

In order to get a global estimate of *V*, a set of marker frequencies from the experiment need to be chosen, and a position for the selected locus relative to these markers must be assumed. Which markers are chosen, and what position is assigned to the selected locus will influence the value of $\hat{V}$. This is because the value of *V* gives a measure of the stochasticity in marker frequency. However, estimating the stochasticity in frequency for any given set of markers, will depend on what is assumed about the location of the selected locus. If a set of markers are chosen in a null region, but assigned as being in a selected region, then the value of $\hat{V}$ may be much higher than what *V* truly is. The same

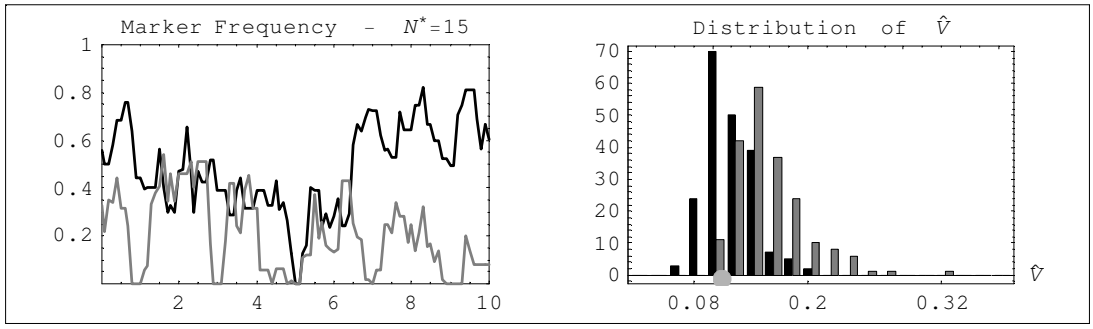
would be true if a set of markers were chosen in a selected region, but assigned as unlinked. Hence, choosing the wrong model for any given set of markers may give an overestimate for the value of *V*, which may adversely affect mapping accuracy, by giving very different settings for the significance levels and confidence intervals. Therefore, in order to estimate *V*, a reasonable assumption must be made about the genetic background of the markers that are used for the estimation.

One reasonable assumption that could be made is that most markers in the genome would either be unlinked or loosely linked to a selected locus. Therefore $\hat{V}$ could be calculated using all markers on the genome, and assuming they are all unlinked. In this case $\hat{V}$, which is a function of *x* the location of the selected locus, can be calculated by setting $x = \infty$. Any deviations of this estimate from the true value of *V* will mainly be caused by markers that are closely linked to selected loci that are at extreme frequencies. So, by using this method to estimate *V*, the accuracy of the estimation is expected to decrease with an increase in the number of selected loci at extreme frequencies.

An example of this is shown in Figure 4.1. It shows the typical marker frequency in a selected population for two different effective initial population sizes *N**, different numbers of selected loci, and the corresponding $\hat{V}$. In Figure 4.1(a) $N^* = 15$. The marker frequency graph shows the typical marker frequency when there is just a single selected locus (black curve, selected locus at position 5), and the typical marker frequency when there are five selected loci (grey curve, selected loci at position 1, 3, 5, 7, 9). The corresponding bar charts in Figure 4.1(a) represent the distribution of $\hat{V}$ for the single selected and five selected loci examples over 200 replicates. The filled gray circle in the bar charts represents the true value of *V*. The same setup is depicted in Figure 4.1(b) with $N^* = 150$. In both bar charts it can be seen that $\hat{V}$ is generally overestimated, with the overestimate being wider for multiple selected loci. Also with $N^* = 150$, the overestimation is much wider when compared to $N^* = 15$. This is because when *N** is large, there is less of an overlap between the distribution of

frequency between null and linked makers, and therefore the linked markers have a more detrimental effect on the estimate of V when N^* is large. Overall, however, it can be seen that when N^* is small the estimate will provide a reasonable approximation to the true value of V, for both the single and multiple selected loci case. When N^* is large the estimate is a lot less accurate, but in practice this large inaccuracy will not pose too much of a problem. This is because, as discussed in Chapter 3, it would be highly unlikely we would see the pattern of marker frequency as shown for the multiple selected loci case in Figure 4.1(b). If there were many selected loci fixed in the selected population, the marker frequency pattern will most likely resemble the pattern in Figure 4.1(a), and hence the estimate of V will not be that inaccurate. Secondly, it will be shown later that when N^* is large the mapping accuracy is very good. Consequently, for large N^* the estimate of V need not be very accurate. That is, when N^* is large, the estimate of V could be a very large overestimate, but it will most likely have no adverse effect on the overall mapping accuracy.

(a)



(b)

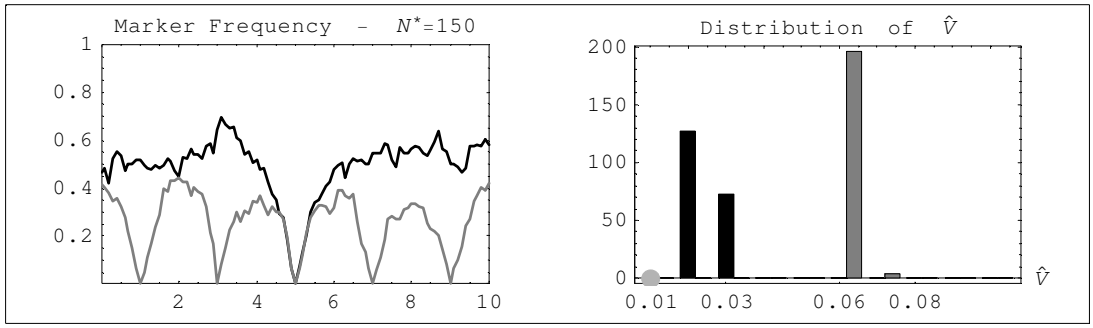


Figure 4.1 – (a) the first graph plots the typical marker frequency when N^*=15. The black curve is the marker frequency when there is a single selected allele fixed at position 5. The gray curve is the marker frequency when there are five selected alleles fixed at positions 1, 3, 5, 7, 9. The bar chart plots the distribution of $\hat{V}$ for the two marker frequency plots. The gray filled circle represents the true value of V. (b) plots the same information as (a) except N^*=150.

4.2.5 Estimating location *x*

Once the global estimate of V has been obtained, it will be used in the calculation for each likelihood ratio. As a result, the one unknown parameter left in the likelihood ratios is x, the location of the selected locus between the interval being tested. For the null hypothesis, it is assumed that the markers are not linked to the selected locus. Hence setting $x = \infty$ in (4.1) gives the maximum likelihood for the null hypothesis. For the alternative hypothesis, an estimate of x is needed. In the alternate hypothesis, we are testing for the presence of a selected locus over specific intervals. Hence, (4.1) needs to be maximized over an interval $\theta = (x_1, x_2)$, where x_1 and x_2 are the boundaries of the specific interval that is being tested. Since the distance between x_1 and x_2 is usually relatively small, typically 10-20 cM, the easiest way to find this maximum likelihood estimator is to sample various values of x over the interval θ, say every 0.5 cM, and choose the value of x that maximizes $l(x, \hat{V})$. Therefore, for each interval that is tested, the log likelihood ratio λ can be defined as $\lambda = l(\infty, \hat{V}) - \max_{x \in \theta \cup \infty} l(x, \hat{V})$.

4.2.6 Significance Levels

Once the log likelihood ratio λ has been calculated for each interval, the significance of the ratios needs to be tested. The smaller the value of λ, the less likely it is that the data would be seen under the null hypothesis. How small λ should be can be determined by obtaining a genome wide significance level. This significance level can be obtained by

determining the distribution of the minimum value of λ, say λ_{min}, when the marker frequencies come from a null region. This distribution can be obtained by simulation. That is, a joint distribution of marker frequency is simulated from a null region, where the marker setup is identical to the marker setup in the genome in the experiment. From this simulated data, the likelihood analysis is run along each chromosome in the genome, and the value of the minimum likelihood ratio is recorded. This process is repeated a number of times to get the distribution of λ_{min}. From this simulated distribution, a threshold δ can be obtained such that, say $P(\lambda_{\min} \leq \delta) = 0.05$. Therefore any likelihood ratio below δ can be deemed to be significant.

4.2.7 Confidence Intervals

In order to obtain a confidence interval for the predicted location of the selected locus, the distribution of the distance between the predicted location of the selected locus and true location is needed. This distribution can be obtained by simulation. That is, data are simulated from the hypothesis that a selected allele is fixed in the population. This can be done by simulating the joint distribution of marker frequencies linked to a selected allele that is fixed in the population, where again the marker setup is similar to the marker setup in the experiment. From these simulated data, the likelihood analysis is applied along the regions linked to the selected locus (say, ± 50cM from the selected locus). If the log likelihood ratios around the selected locus are below the significance level, then the distance between the position predicted by lowest significant log likelihood ratio around the selected locus and the position of the selected locus is recorded. This process is repeated to get a distribution for the distance. Using this distribution, a confidence interval can be obtained for the predicted location of a selected locus.

4.2.8 Simulating Data

The data needed for constructing the significance levels and confidence intervals can be obtained directly by simulating data from a multivariate Gaussian, or by simulating a branching process and using the resulting marker frequencies. For the genome wide significance level, simulated data is needed from a null region. The Gaussian approximation provides a very good fit for markers in a null region, and the therefore the simulated data needed for the significance level can be obtained directly by simulating frequencies from a multivariate Gaussian with parameter $\hat{V}$. When determining confidence intervals, simulated data are needed from markers that are linked to the selected locus. In this case, a branching process simulation may provide a more accurate fit for the distribution of frequency. This is particularly true for markers that are very tightly linked to the selected locus and/or when the initial population size is very small. To simulate a branching process, parameters n, μ, σ^2 and t need to be obtained from $V = 1/n(1+(\sigma^2(1-\mu^{-t})/\mu(\mu-1)))$. n is the expected number of initial recombinants with the selected allele, μ is the mean offspring number per generation, σ^2 is the variance in offspring number per generation, and t is the number of generations of selection. Therefore to simulate the branching process, in each replicate N recombinant individuals are generated, all of which contain the selected allele and selected for t generations, where $N = Bin(2n,0.5)$ is a binomially distributed random variable. The offspring distribution of each of the N recombinant individual is not important, as it is assumed in the model that the initial population size is not extremely small, and consequently most marker frequencies can be reasonably approximated by a Gaussian. Hence, for simplicity it can be assumed that the offspring distribution per generation is Poisson distributed, and consequently $\sigma^2 = \mu$. Therefore, assuming the value of t is known, values for n and μ can be obtained by solving $\hat{V} = (\mu + \mu^{-t})/n(\mu - 1)$, and ensuring n is not too small (ie. say, $n \geq 10$).

4.2.9 Mapping Accuracy – False Negative Rate

The accuracy of this method in detecting selected loci depends mainly on the size of the effective initial population N^* at the selected generation. When N^* is low, there can be large stochasticity in marker frequency in unlinked regions. As a result, the ability to statistically distinguish selected regions from null regions diminishes when the effective initial population size is low. An example of this is shown in Figure 4.2. Figures 4.2(a), 4.2(b) and 4.2(c) show the likelihood analysis applied to an example when $N^* = 30$, $N^* = 20$ and $N^* = 13$. The graphs plot the log marker frequencies (thin solid black line), the deterministic expectation (dotted black line), the log likelihood ratio (thick gray line), and the significance level (thick solid black line). All three examples have a single selected allele fixed at position 1 and a marker every 5cM. To calculate the log likelihood ratio, the genome was split into overlapping intervals of 10cM, where the overlap was 5cM. For each interval, the log likelihood ratio was calculated using the two markers that define the interval. When $N^* = 30$ there is one clear drop in the log likelihood ratio which highlights the location of the selected allele. The same is true when $N^* = 20$, but there is slightly more variation in the marker frequency, and hence the significance level is moved lower. When $N^* = 13$ there is a lot more variation in marker frequency in unlinked regions, and as a result, more valleys in the log likelihood map appear in unlinked regions. Consequently, the significance level is set to a stricter level to eliminate the spurious signals in these unlinked regions. However, this also results in the true signal at position 1 being discarded, as it does not exceed the significance threshold, resulting in a false negative. Figure 4.2(d) plots the false negative rate against N^*. It can be seen that the false negative rate is 0% for most N^* and only starts to increase for very small N^*. So overall, given that we have exact marker frequencies, this method will only have difficulty in identifying the location of fixed selected alleles when N^* is extremely small.

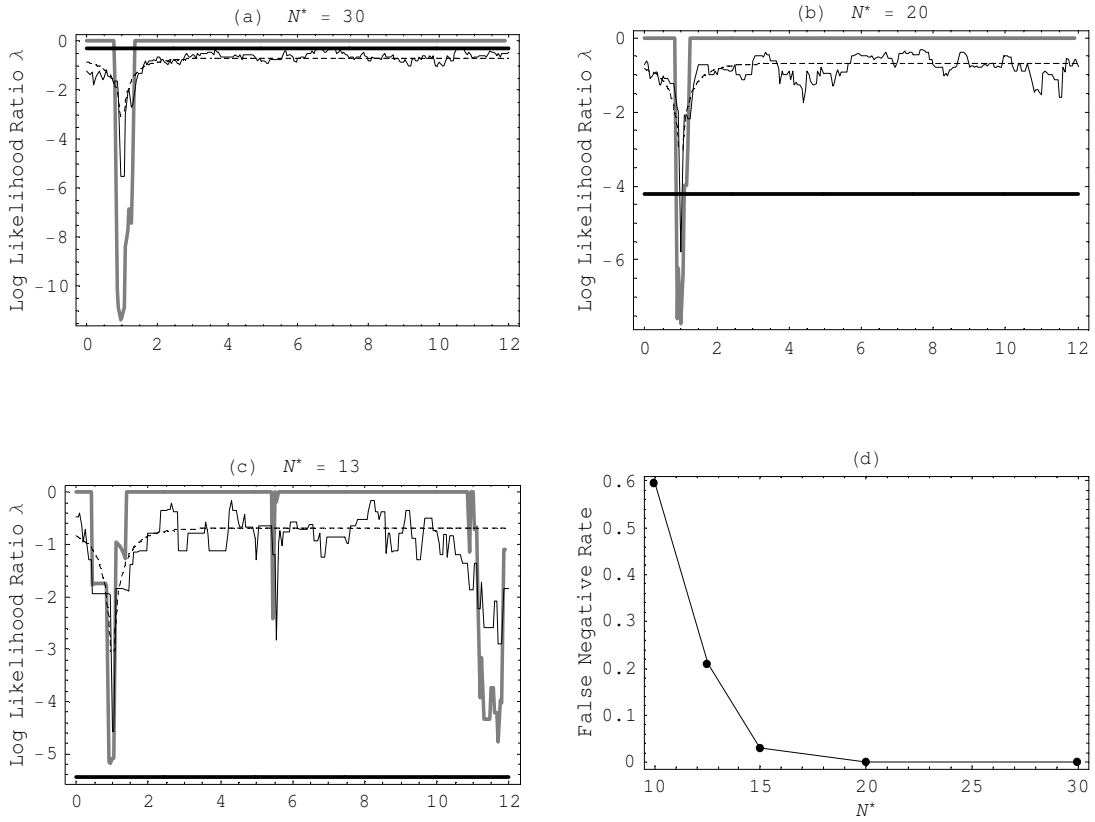


Figure 4.2 – (a), (b) and (c) plot of the log marker frequencies (thin solid black line), the deterministic expectation (dotted black line), the log likelihood ratio (thick gray line), and the significance level (thick solid black line) when the effective initial population size N^* is 30, 20 and 13 in the selected generation. In all three graphs there is a single selected locus at position 1. (c) plots the false negative rate against N^*.

4.2.10 Number of Flanking Markers

In the examples in Figure 4.2, the log likelihood ratio λ for each interval was calculated using just the two markers that defined the interval. It may be possible to improve the false negative rate for very small N^* by increasing the number of flanking markers used in the calculation of λ. However, it turns out that increasing the number of flanking markers for the calculation of λ at each interval, actually increases the false negative

rate. Figure 4.3(a) illustrates this. The dotted curve in Figure 4.3(a) plots the false negative rate when there were six markers used in the calculation of λ for each interval. That is, three markers on each side of the interval with a spacing of 2cM between each marker. The solid black line is the false negative rate when only two markers are used (ie. the same graph as Figure 4.2(d)). From this graph, it can be seen that for large N^*, increasing the number of flanking markers makes no difference to the false negative rate, as the false negative rate remains at zero. For the smaller N^*, the false negative rate increases. So, increasing the number of flanking markers actually decreases the mapping accuracy for small N^*.

The main reason why this happens is because there can be very sharp changes in frequency between tightly linked markers. This causes very extreme likelihood ratios to be present in unlinked regions. As a result the significance level has to be set much stricter in order to avoid false positives, but this also eliminates a lot of true signals. An example is shown below in Figure 4.3(b) and 4.3(c). Figure 4.3(b) shows the marker frequency in a single replicate when N^*=15. There is a single fully selected locus at position 1. Figure 4.3(c) plots the corresponding log likelihood ratio when there was just a single flanking marker used for the calculation of λ for each interval (gray curve), and also a plot of the log likelihood ratio when there were 3 flanking markers used for the calculation of λ for each interval (dotted curve), where the spacing between each flanking marker was 2cM. In Figure 4.3(c), there are two drops in the log likelihood ratio map. The first drop is near position 1, which is the location of the selected locus, and the second drop, around position 2.7, is in a null region. For the drop in the null region, the log likelihood ratio associated with the 3 flanking marker setup is much more extreme than the single flanking marker setup. This is because, at that position 2.7 in the null region, there is a sharp change in frequency in the flanking markers used for that particular interval. This sharp change in frequency between the tightly linked markers in the flanking regions, results in the log likelihood ratio being much more extreme. As a result the significance level has to be set much more strictly, resulting in a much higher false negative rate.

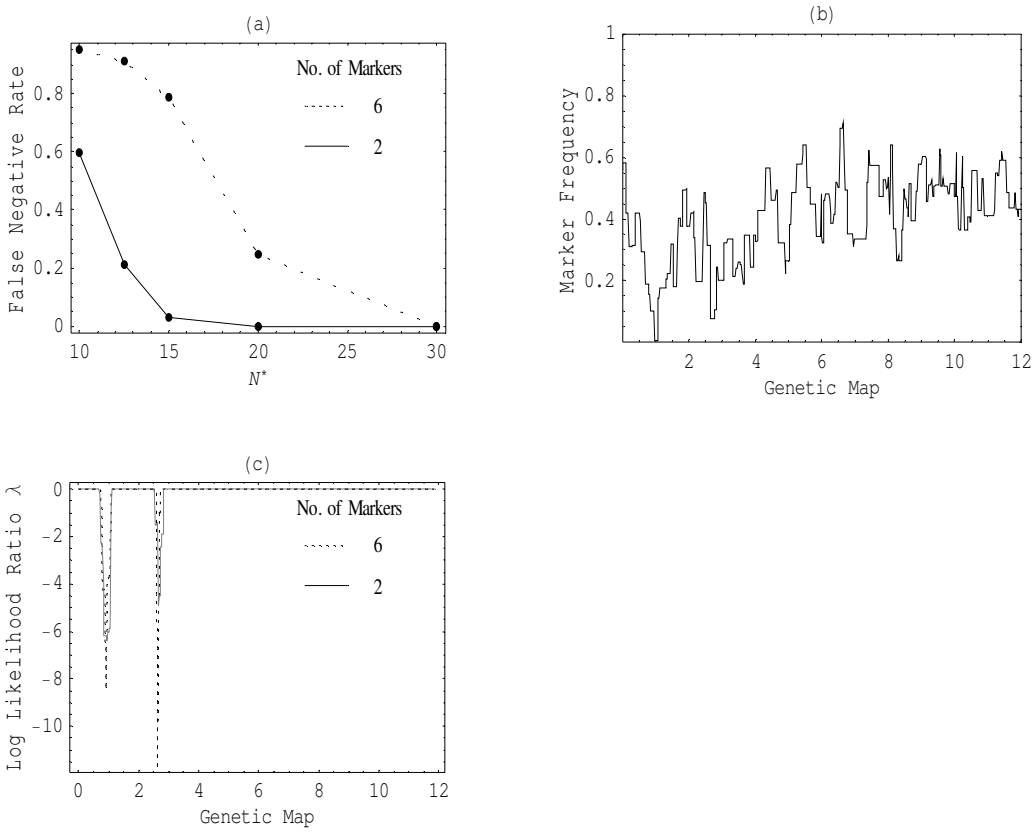


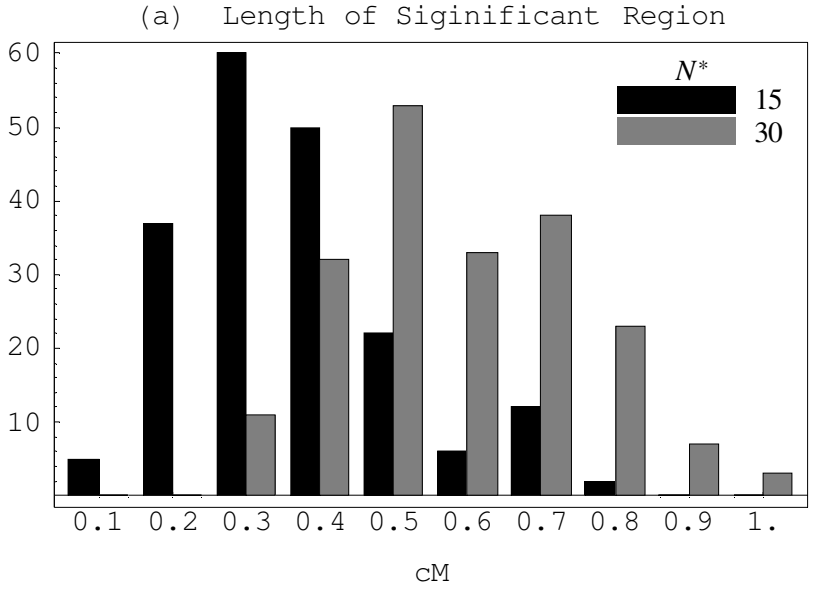
Figure 4.3 – (a) plot of the false negative rate against N^*. The solid line is the false negative rate when only two markers were used in the analysis of each interval (ie. the two markers that define the interval). The dotted line is the false negative rate when six markers were used in the analysis of each interval (three markers on each side of the interval, with a spacing of 2cM between each marker). (b) plot of the typical marker frequency seen in a single replicate when N^*=15. There is a marker every 2cM, and there is a single selected allele fixed at position 1. (c) plot of the log likelihood ratios for the example shown in (b). The solid line is the log likelihood ratio when two markers were used in the analysis of each interval. The dotted line is the log likelihood ratio when six markers were used in the analysis of each interval.

4.2.11 Mapping Accuracy – QTL Location

Given that significant changes in marker frequency have been identified, how accurate is the predicted location of the selected locus? The estimate of the location of the selected

locus will be the estimate provided from the lowest likelihood ratio from a region of genome flagged as significant. It can be seen from Figure 4.2 that the size of this region around a selected locus that is flagged as significant differs with *N**. Figure 4.4(a) plots the distribution of the length of the significant region for two examples $N^* = 15$ and $N^* = 30$. It can be seen that with the larger *N** the length of genome that flags as significant around the selected locus can be much larger than when with the smaller *N**. Therefore, for any single selected locus, the signal associated with it can stretch out for a very long distance. Hence, a large region of genome that flags as significant may not necessarily contain many selected loci, but could just be the signature of a single selected locus.

The accuracy of the predicted location also depends on the size of *N**. The larger *N** is, the more accurate the prediction. Figure 4.4(b) plots the length of the 90% confidence interval against *N**. These confidence intervals were calculated from 200 replicate simulations, where genome was split into overlapping intervals of 10cM, where the overlap was 5cM. It can be seen from Figure 4.4(b), that as *N** get smaller the confidence interval of the predicted location gets wider.



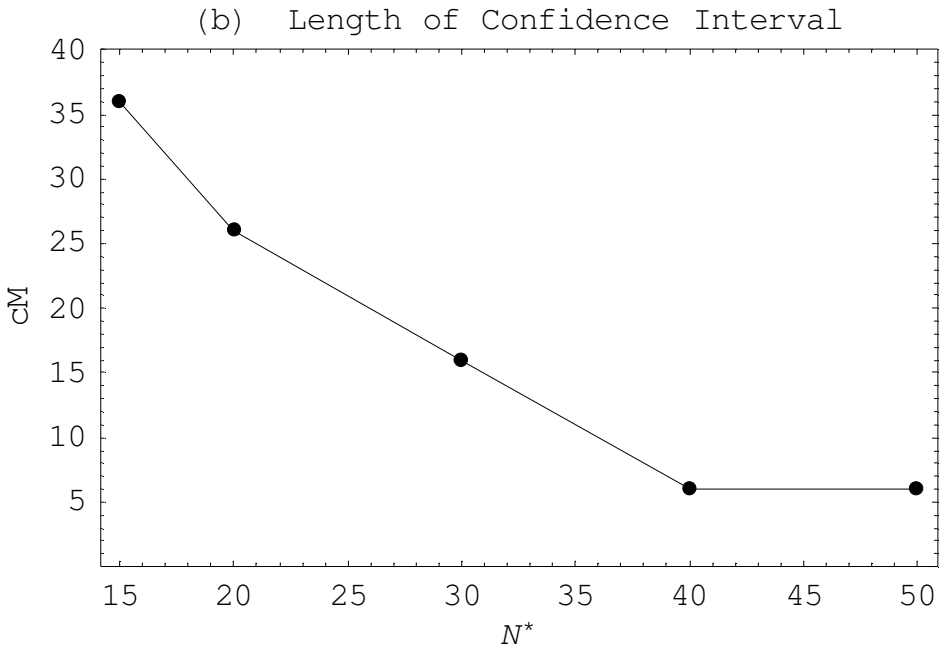


Figure 4.4 – (a) plot of the distribution of the length of the significant region around the selected locus, for N^*=15 and 30. (b) plot of the length of the 90% confidence interval of the predicted location of the selected locus against N^*.

4.3 Parental Genotypes in Population

In the models discussed so far, it was assumed that the initial cross population consisted of only recombinant genotypes. However, in some experimental setups this may not be true and the cross population may be a mixture of both parental and recombinant genotypes. This typically happens in genetic systems where it is difficult to make individual crosses. For example, in malaria parasites, the sexual stage of the life cycle happens inside the gut of a mosquito. Therefore, when a cross is to be made between two strains of malaria parasites, a mixture of the two parental strains is given to the mosquito, and the parental strains randomly cross with each either inside the mosquito gut. As a result it is very difficult to control which parental genotypes cross with each other. Therefore, due to this random crossing between the strains, genotypes of the same

parental strain will cross. Hence, the resultant asexual cross progeny will not consist solely of recombinant genotypes, but will also contain parental genotypes.

4.3.1 Likelihood Model with Parentals

In the following model it is assumed that the initial population consists of both asexual recombinant genotypes and asexual parental genotypes. Similar to a recombinant model, this asexual population is selected for a trait for t generations, and marker frequencies are analysed in the selected population. From this model, the aim is again to concentrate on the simplest case and search for selected alleles that have fixed in the population. If a selected allele has fixed in the population, then only one of the two original parental genotypes will be present in the selected population, along with the surviving recombinant genotypes. The distribution of marker frequency in such a population, can be described by the distribution of marker frequency obtained from a population consisting of a single fitness class, which is selected for t generations, where the initial population consists of a proportion p_r of recombinant genotypes, all of which contain a single selected allele, and a proportion 1-p_r of the fitter parental genotype. This distribution can once again be approximated by a multivariate Gaussian, with the log likelihood function given by (4.1) with slight modifications made to the moments. The mean vector in (4.1) is now given by $m_i = p_r P_i$, and the covariance matrix $\Sigma_{ij} = V * C_{ij}$ is modified such that $C_{ij} = p_r(P_{ij} - p_r P_i P_j)$. Here, P_i is the probability that marker i is on a recombinant genotype in the initial population, and P_{ij} is the probability that marker i and marker j are on a recombinant genotype in the initial population.

Using these moments with the Gaussian, the likelihood method is applied in the same way as the recombinant model. The one difference now is that there is an extra unknown parameter p_r in the model. This parameter p_r is a global parameter and is only needed to be estimated once. To estimate p_r the same methodology and logic are used as in the

estimation of V. That is, to estimate p_r, all markers from the experiment will be used for the estimation, and it will again be assumed that all these markers are unlinked. Substituting the maximum likelihood estimator $\hat{V} = ((y-m)'C^{-1}(y-m))/k$ (which is now a function of p_r) into (4.1), we get that the log likelihood function of the multivariate Gaussian in given by $-0.5(n + nLog(\hat{V}) + Log(\det(C)))$ which is a function of p_r. Maximizing this function with respect to of p_r gives the maximum likelihood estimator $\hat{p}_r$.

$$\hat{p}_r = \max_{p_r \in (0,1]} (-0.5(n + nLog(\hat{V}) + Log(\det(C)))) \quad \textbf{(4.2)}$$

The easiest way to obtain $\hat{p}_r$ from (4.2) is to sample various values ($0 < p_r \leq 1$) and choose the one that maximizes (4.2). The resultant global estimate $\hat{p}_r$ can be used to get a value for $\hat{V}$, and also be used in the calculation of the log likelihood ratio λ for each interval. Therefore, the log likelihood ratio for each interval for this parental model can be defined as $\lambda = l(\infty, \hat{V}, \hat{p}_r) - \max_{x \in \theta \cup \infty} l(x, \hat{V}, \hat{p}_r)$

4.3.2 Malaria Data

In this section, the likelihood model is applied to some data obtained from an experiment aimed at mapping genes responsible for drug resistance in malaria parasites. Figure 4.5(a) plots the data obtained from a single replicate of such an experiment. It shows the marker frequencies across 14 chromosomes. There are between 2 and 10 markers on each chromosome, with an average distance of 10cM between each marker.

In order to analyse this data, some adjustments must be made to the current model to reflect that this data does not consist of true marker frequencies, but rather estimates of the true frequencies. Therefore, the distribution of the error rate of the marker frequency estimation must be incorporated into the model. So, let the observed frequency $p^* = p + \varepsilon$, where p is the true marker frequency and ε is the frequency measurement

error. For simplicity, I will assume that the measurement error ε for each marker is Gaussian distributed. This is not ideal, as when p is near the extremes, adding a Gaussian ε may result in p^* going beyond the frequency range of 0 – 1 in the model. However, this will only happen in rare cases. So, assuming ε is Gaussian distributed, p^* will also be Gaussian distributed. In order obtain the moments of p^* the moments of ε are needed. For this dataset, the measurement error ε is slightly different for each marker. However, from separate experiments calibrating the marker frequency estimates and marker counts within pools, it was found that in general $E(\varepsilon) = \pm 0.05$ for most markers in the dataset. Therefore, as specific measurements for $E(\varepsilon)$ are not available for the markers, a uniformly distributed random number between -0.05 and +0.05 is assigned to $E(\varepsilon)$ for each marker. The $Var(\varepsilon)$ can be obtained directly from the data, as for each marker several measurements of the frequency were recorded. Therefore, for each marker the variance of these repeated measurements is used as $Var(\varepsilon)$. Also, when specific information about the distribution of ε for each marker is not available, it is best to remove very tightly linked markers from the analysis. This is because the true frequencies of very tightly linked markers are highly correlated. However, when only vague information about ε is input into the model for each marker, this high correlation in frequency between tightly linked markers may not be captured by the model. This may result in the model predicting far more variance than is present in the selected population, which will greatly reduce the mapping ability. Hence, for this dataset, as precise information about the distribution of ε is not available, markers that are spaced less than 4cM apart have been excluded from the analysis.

With these adjustments made to the model, the first step needed to calculate the log likelihood ratios for this data is to get a value for $\hat{p}_r$, the estimate of the proportion of recombinants in the effective initial population. Using (4.2) we get $\hat{p}_r = 1$. This indicates that there are no parental genotypes in the selected population. In practice, if there were parental genotypes present in the initial population, and a selected allele has fixed (or nearly fixed) in the selected population, then the most likely reason no parental

genotypes remain in the selected population, is that the fitter parental genotype was at a low proportion in the initial population, and/or the recombinant genotypes have a higher fitness than the parental genotypes. Either of these scenarios would result in a low or negligible proportion of the fitter parental genotype in the selected population. With this particular dataset, it is most likely the fitter parental genotypes were present at a low proportion in the initial population, as opposed to recombinant genotypes having a higher fitness. This is mainly because if recombinant genotypes had a higher fitness, and negligible parental genotypes remain in the selected population, then we would expect more than one selected allele to be nearing fixation. However, from Figure 4.5(a), we see there only seems to be a single locus nearing fixation.

With the value of $\hat{p}_r$ obtained, the next step is to get a value for $\hat{V}$. The exact value of $\hat{V}$ will depend on what values from the uniform distribution were assigned to $E(\varepsilon)$. Figure 4.5(b) shows the distribution of $\hat{V}$ over a 1000 replicates. In each replicate $E(\varepsilon)$ was assigned a different value for each marker. From Figure 4.5(b), $\hat{V}$ ranges from 0.063 to 0.112, with an average of 0.084. The median value, call it $\overline{\hat{V}}$, from this distribution and the corresponding values of $E(\varepsilon)$ will be used in the calculation for the log likelihood ratios.

Using the estimate $\hat{p}_r$ and $\overline{\hat{V}}$, the log likelihood ratio λ was calculated between each consecutive pair of markers on each chromosome. Once this was complete, the genome wide significance level was obtained. The simulated data needed for constructing the genome wide significance level was obtained by simulating marker frequencies directly from a multivariate Gaussian using the median estimate $\overline{\hat{V}}$. The marker setup, and the moments for ε, in the multivariate Gaussian was exactly the same as in the experimental data. Figure 4.5(c) plots these log likelihood ratios λ (thick gray line) and the genome wide significance level (black line). From this Figure, there is a single location on chromosome 7 which is predicted to have a selected allele that has fixed. The estimated

location is 2 cM from the second last pair of markers on that chromosome. It should be noted that the estimate of the location of the selected allele in the interval will always be biased towards the edges of the interval. This is because the Gaussian approximation is not a good fit to markers that are very tightly linked to selected alleles that are fixed. So, the likelihoods tend to have a higher value when testing for fixed selected alleles very close to the markers in the interval than in the middle of the interval.

The 90% confidence interval was ±13 cM from that location. The simulated data needed to derive this confidence interval was obtained by simulating a branching process using the parameters $n = 18$, $\mu = 3$ and $t = 10$. These parameters n and μ were obtained by solving $\overline{\hat{V}} = 0.084 = (\mu + \mu^{-t})/n(\mu - 1)$. For the branching process $N = Bin(2n, 0.5)$ recombinant individuals, all of which contain the favorable allele, are selected for $t = 10$ generations. The marker setup in the branching process was broadly the same as the experiment, in that markers were spaced every 10cM. The selected allele was positioned between the middle of two markers. Finally, in each replicate, for each marker a Gaussian distributed error is added to the frequency in the selected generation to simulate the error in the observed frequencies. The mean of the Gaussian error was assigned a uniformly distributed number between -0.05 and +0.05, and the variance of the error for each marker was assigned the mean variance from the experimental data. However, for confidence intervals, since the simulated frequencies would be near the extremes, some frequencies may go beyond 0 or 1 when adding this Gaussian error. Therefore, checks were done for this and simulated frequencies that went beyond 0 or 1 were removed, and a new Gaussian error was added.

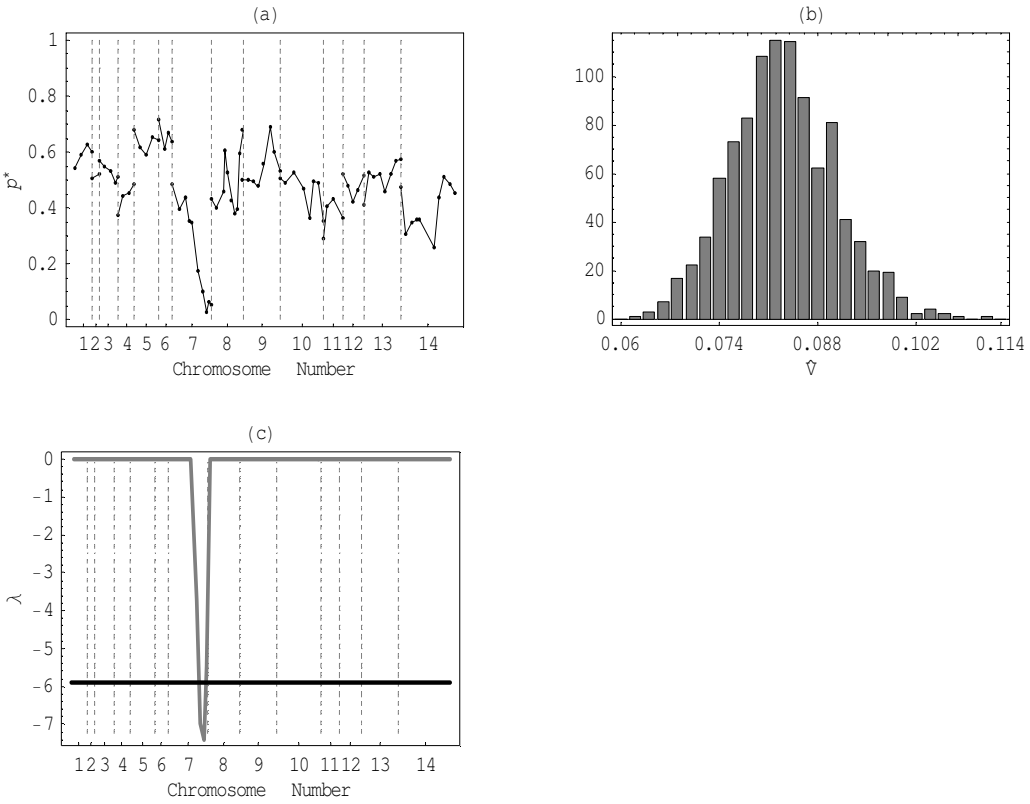


Figure 4.5 – (a) plot of the marker frequency estimates across the 14 chromosomes. (b) plot of the distribution of $\hat{V}$ when different error rates are assigned to each marker. (c) plot of the log likelihood ratios (gray curve) and significance level (black line) for the data shown in (a).

4.4 Discussion

The main aim was to develop a statistical model to search for selected alleles that have fixed in the selected population. It was shown in Figure 4.2 that if a selected allele has fixed in the population, then the statistical model has a high probability of identifying the location of the selected locus, as long as the effective initial population size N^* at the selected generation is not extremely small. It was also shown that analyzing just two markers at a time in the likelihood model gives optimal mapping ability. This suggests that having a very dense map of markers is unnecessary for this analysis. Typically, as

the length of an interval being analysed along a chromosome would be about 10cM, a reasonable approach for the total number of markers to use on each chromosome could be to use no more than a marker every 10cM.

All results in this chapter were based on analyzing true marker frequencies from an experiment. However, as seen with the malaria experiments, the data may not consist of true marker frequencies, but estimates of the true frequency. If this measurement error ε between the estimated frequencies and true marker frequencies is large, then the mapping ability may decrease, particularly for small N^*. The amount of decline in the mapping ability will depend on the distribution of ε, and also the amount of information that is available about ε for each marker. Figure 4.6(a) shows a simple example of this. It plots the false negative rate when $N^* = 50$ for different error rates. The marker frequencies used to calculate the false negative rates were simulated from a branching process, with a Gaussian distributed error added to each marker frequency in the selected population. For the Gaussian error, $Var(\varepsilon) = 0.001$ for each marker (this was roughly mean variance in the experimental data) and $E(\varepsilon) = U(-q, q)$ for each marker, where U is a uniform distributed number between $\pm q$. The false negative rates in Figure 4.6(a) are plotted against different values of q. The solid curve in Figure 4.6(a) plots the false negative rate when only general information about $E(\varepsilon)$ was input into the likelihood model. That is, in each replicate, for each marker, the true value of $E(\varepsilon)$ is not assigned to each marker, but instead a uniform distributed number between $\pm q$ is assigned to $E(\varepsilon)$. The dotted curve in Figure 4.6(a) plots the false negative rate when the true value of $E(\varepsilon)$ was used for each marker in the likelihood model. The results show that, as expected, the larger the error ε, the greater the false negative rate. Although in this example the false negative rate only starts to increase when the error rate gets relatively large, when N^* is lower, the false negative rate would increase for smaller error rates. Figure 4.6(a) also gives an indication of how precise the information about ε needs to be. It can be seen that when $E(\varepsilon)$ is not too large, precise information about it is not necessary. Inputting very general information about $E(\varepsilon)$ into likelihood model provides as good accuracy as inputting the correct value of $E(\varepsilon)$. However, when the

error rate gets large having accurate information about $E(\varepsilon)$ for each of the markers reduces the false negative rate. So, overall, given N^* is not too small, and ε is not very large, the error rate in the frequency estimation should not pose a problem if searching for fixed selected alleles, provided that very rough information about the distribution of ε are available.

Although the model developed in this chapter specifically searched for selected alleles that had fixed in the population, it should also be able to identify selected alleles that have not yet fixed, but generally increasing in frequency and are nearing fixation. Figure 4.6(b) plots the false negative rate against the mean frequency of the selected allele. The dotted line plots the false negative rate when $N^* = 50$ and the solid line plots the false negative rate when $N^* = 150$. It can be seen in both graphs that the false negative rate starts to decline when the mean frequency of the selected allele is less than 0.9. When the mean frequency is 0.8 there is a very small probability that the model is going to identify the selected locus. The graph with the smaller N^* has slightly better accuracy as there is more variation in frequency in the linked regions than the higher N^*.

So, given that the present model will only be able to detect very high frequency selected alleles, a separate model must be used if the aim is to search for selected alleles at lower frequencies. The specific model that would be needed would be the distribution of marker frequency in a selected population where there are two fitness classes remaining. This model will specifically search for selected alleles that are not fixed. However, in practice, it may not be easy to apply this more complex model to data. Firstly, as a result of an extra fitness class being included in the model, more unknown parameters will exist in the model. Estimating these unknown parameters from the data will increase the computation time for the likelihood analysis. Secondly, when selected alleles are not fixed in the population, there can be a very big overlap between the distribution of frequency of the linked and unlinked markers. Hence, very accurate measurements of the marker frequencies would be needed in order to statistically distinguish linked and unlinked markers. Therefore, if true marker frequencies are not attainable and only

vague information about the frequency measurement error of each marker is available, then it will be very difficult for this branching process model to identify true selected alleles.

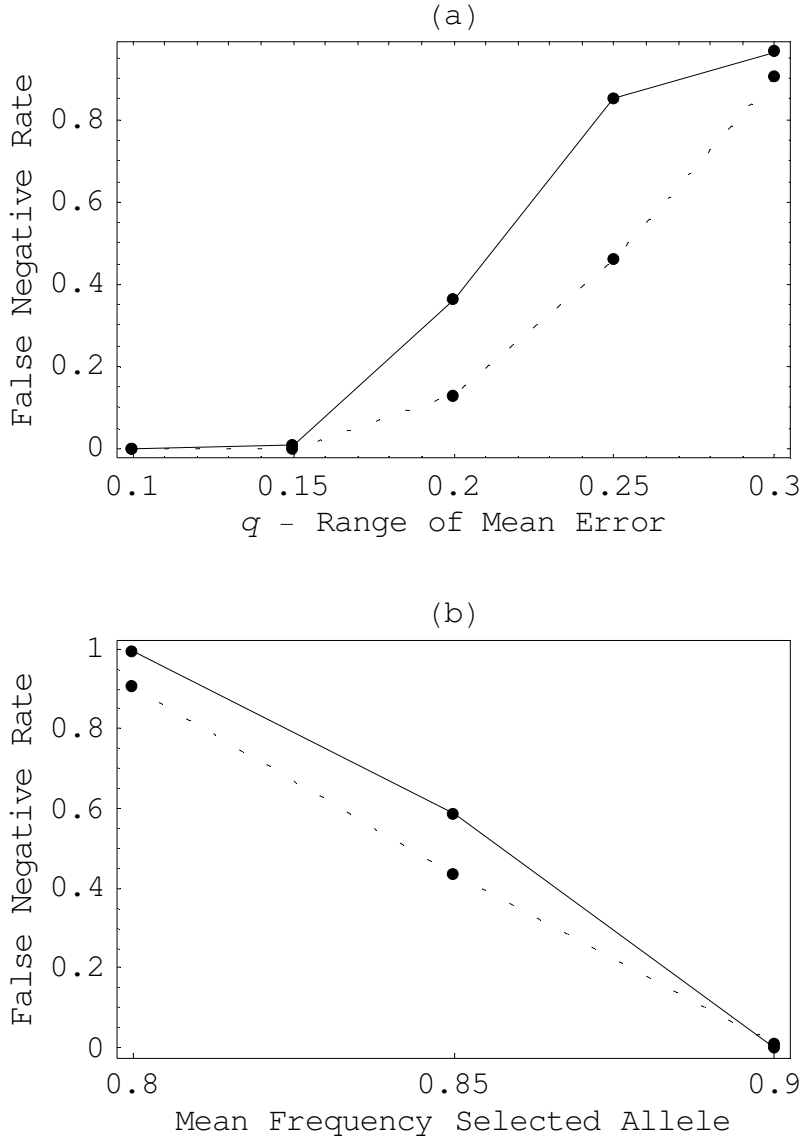


Figure 4.6 – (a) plots the false negative rate against varying error rates in the marker frequency estimation. The dotted line plots the false negative rate when exact details of the distribution of the error rate for each marker was used in the likelihood model. The solid line plots the false negative rate when only vague details of the distribution of the error rate for each marker was used in the likelihood model. (b) plots the false negative rate when the selected allele has not yet fixed in the selected population. The dotted line is the result when N^*=50 and the solid line is the result when N^*=150.

Chapter 5: Discussion

This thesis examined the strategy of identifying loci that influence a trait by analysing marker allele frequencies in pools of asexual cross progeny of extreme phenotype. These pools of extreme individuals were selected by multiple generations of asexual reproduction and selection. The two main topics I analysed were, how effective this method is in identifying loci for different types of genetic models, and how to statistically analyse marker frequencies in the selected population using a branching process model.

In terms of the effectiveness of this method, I showed that the ability to identify markers linked to a causative locus depends on the variation in marker frequency in the selected population. The larger the variation in marker frequency, the more chance there is of spurious peaks and valleys in frequency in unlinked regions. The amount of variation in frequency in unlinked regions will be determined by the number of unique recombinant genotypes present in the selected population. The more unique recombinant genotypes present in the selected population, the more balanced the representation of markers in the selected population, and the more likely that the marker frequency is to approach the deterministic expectation, making identification of causative loci much easier. The number of selected loci, the size of the initial population, and the length of time of selection, will determine how many unique recombinant genotypes are left in the selected population.

Specifically, I showed that for Mendelian traits, the initial population size need not be that large. Ideally selection should be continued until the fitter class of recombinants have fixed (or nearly fixed) in the selected population. In this case, since half the initial cross would be expected to belong to the fittest class, large numbers of unique recombinant genotypes should be present in the selected population. The precise size of

the initial cross needed in order to minimize the variance in frequency will depend on the variation in the number of descendants each of the fitter initial recombinants leave in the selected population. The smaller this variation, the smaller the initial cross needs to be. Overall, I showed that having an initial population size at least in the low to mid hundreds should ensure that there is only a small probability of spurious changes in frequency in unlinked regions.

For quantitative traits, however, I showed that the initial population size needs to be much larger if the aim is to select out the most extreme class of recombinants. However, I also showed that, unlike the single selected locus model, selecting out only the most extreme individuals may not be possible or even desirable. It may not be possible, as when a large number of loci influence the trait, the fittest possible genotype may not have been produced at meiosis or may have simply been lost in the initial few generations due to low numbers. Hence, if selection is continued for long enough the alleles that fix at the selected loci may not be the fittest. This issue is not specific to just this experiment, but highlights a general inefficiency of selection in asexual populations in that selection alone will find it very difficult to pick out the fittest possible genotype, and is consequently one of the hypotheses for the evolution of sex and recombination. That is, if reproduction was sexual, then recombination can shuffle the genotypes in each generation. This shuffling process enables fitter genotypes to be produced much more quickly, as there is a greater probability that the beneficial alleles that are initially spread across various individuals in the population can be brought together on to a single recombinant genotype (Crow & Kimura, 1965; Rice, 2002). However, with only one generation of recombination present in this experiment, selection is much less efficient and can only choose from the genotypes that exist in the initial cross. So, with the fittest genotype unlikely to be produced at meiosis, one has to rely on mutations that arise during the selection phase to generate the fitter genotypes that were not initially present or lost in the initial few generations. However, accumulating beneficial mutations can be quite difficult in an asexual population due to clonal interference, which is the competition between beneficial mutations that may arise on separate genotypes. It means

that in asexual populations, unless beneficial mutations arise on the same genetic background, multiple beneficial mutations cannot establish in the population simultaneously. The mutations must be fixed sequentially. This hinders the probability that the fittest possible genotype will be generated and selected. So, overall if the goal was to select out the fittest possible genotype, and there are a large number of loci affecting the trait, then it is very unlikely that this is going to be achieved with the current experimental setup

So, if selection is continued until fixation, the genotype that fixes would be the fittest genotype that survives the initial few generations. However, I showed that unless very large initial population sizes are available, selecting until a single fitness class fixes may not be desirable. This is because the longer selection is applied the more variation in marker frequency would appear in unlinked regions. This is because in any genetic model the initial population will consist of a distribution of fitness classes. This will range from just two fitness classes in a one selected locus model to 2^n fitness classes for an n unlinked selected loci model. In all these genetic models, as selection is applied the less fit genotypes will be lost, and the selected population will increasingly become biased towards descendants of just the fitter initial recombinants. Hence, the number of unique recombinant genotypes in the selected population decreases as selection is applied, and as a result there will be an increase in the variation in frequency in unlinked regions. Therefore, if a very large number of loci influence the value of the trait, and selection is continued for a very long time, there could be very large stochasticity in marker frequency in unlinked regions, as only the fitter initial recombinant descendants would have survived. Since, these fitter initial recombinants would most likely have been at low numbers in the initial population, many spurious peaks and valleys in marker frequency would appear in unlinked regions.

Consequently, finding an optimal length of time to select for would be appropriate for most quantitative traits. However, I showed that finding optimal selection times is a difficult process. In order to find precise selection times that would maximize the ability

to detect selected loci, information about the total number of selected loci and their effects are needed. Obviously, since this is the information we are seeking to obtain from these experiments, this information is not available. Without such information, only very general selection times can be obtained. These general selection times can be obtained by deciding on a minimum effect locus that we would like to detect and assuming a certain number of higher effect loci. However, I showed that these general selection times can give very misleading results, and as a result a lot of false positives, if the actual number of higher effect loci is very different to what is assumed. So overall, if selection is going to be continued for a very long time, the best solution is just to have very large initial population sizes. The moment calculations in Chapter 2 and Chapter 3 can be used to get an idea of how large the initial population size needs to be for various numbers of selected loci and selection times.

The other issue I examined with the model of multiple selected loci is the role of epistasis. In sexual populations, selection has been shown to be inefficient in the presence of epistasis, as it can hinder the generation of optimal genotypes. This is because interaction between alleles at selected loci may lead to multiple fitness peaks, which could mean that the population may find it difficult to move to a fitter genotype if a sub-optimal genotype has established. Although this issue is not particularly relevant in these asexual artificial selection experiments, I showed that selection may still be less efficient in the presence of epistasis as selection can become much weaker when many selected loci are interacting. This is because when a very large number of interactions occur between selected loci, it is more likely the fitness difference between alleles at the selected loci will be very small. So, if a large number of interactions occur between selected loci, selection would most likely need to be applied for a much longer time in order to see any appreciable changes in marker frequency at the selected loci. However, once again, if selection is applied for a very long time it is more likely that spurious peaks and valleys will appear in unlinked regions. So, if very large initial population sizes are available and selection is continued on for a long time, then the general marker frequency pattern in the selected population will look no different to that of any additive

model. That is we will still see peaks and valleys in the marker frequency around the selected loci regardless of whether epistasis is present or not. However, I showed that one possible way to detect the presence of epistasis is to measure marker frequencies at several timeframes and identify any large reversals in the direction of marker frequency change. Although this pattern would indicate interaction between selected loci, the absence of such a pattern would not however indicate an additive model.

The final issue I looked at was a statistical model based on the branching process that aims to identify the location of selected alleles by analyzing marker frequencies in the selected population. I concentrated on a statistical model that searches for selected alleles that have fixed in the population. Using simulated data, I showed the model will successfully identify the location of selected alleles that have fixed in the population given that the effective initial population size is not extremely small. However, if selected alleles have not yet fixed in the population, I showed that this particular model is highly unlikely to detect them if the mean frequency of the selected allele is below about 0.85. If the goal is to detect selected alleles at lower frequencies, then this model must be extended to specifically search for non-fixed loci. This, however, is problematic as extra unknown parameters in the model need to be estimated from the data which will add considerably to the computation time. Therefore, this model is only suited to detect large changes in frequency.

The other issue with the statistical model developed in Chapter 4 is that information about the error rates in the marker frequency estimation must be input to the model in order to implement it. I showed in the chapter that when searching for selected alleles that have fixed in the population, only very general information about the error rates are needed, provided that these are not very large. For example, a rough range for the general mean and variance in the error of the estimates would be sufficient. If, however, the error rates are very large, then more precise information of the error rate of each particular marker would be needed or else the mapping accuracy will decrease very rapidly. Also, for more complex models, such as searching for non-fixed selected loci,

very precise information about the error for each particular marker would be needed. This is another reason why this branching process model may not be suited to detecting selected alleles at smaller frequency, as precise information about the error for each marker may be difficult to obtain.

So, given the limited power of this branching process model, it should probably only be used in limited circumstances. For example, if data from only a single selection experiment is available then the branching process method would be appropriate to use. However, if data from several replicates of selected and unselected experiments are available, then methods based on *t*-tests (Ehrenreich *et al.*, 2010; Segre *et al.*, 2006) would be more appropriate. These methods are easier to implement and have a greater ability to detect selected loci than the specific limited model developed in Chapter 4. They are easier to implement as information about the error rates in frequency estimation are not needed, as the observed frequency estimates can be used as the measurements in the *t*-tests. Secondly, there are no lengthy parameter estimations involved. In the algorithm described in Segre *et al.*, (2006) there are no parameter estimations, while in Ehrenreich *et al.*, (2010) there are only simple linear regressions involved. Therefore, there should not be any major computational issues associated with these methods. With regards to mapping power, the *t*-test methods have shown to have a greater ability to detect selected loci, as these methods have successfully identified many selected alleles that have fixed in the population, and also many selected loci at lower frequencies. For example, Segre *et al.*, (2006) demonstrated their method can successfully identify selected alleles with frequencies as low as 70%. Similarly, in Ehrenreich *et al.*, (2010) many of the selected loci detected were not fixed in the selected population and ranged in frequency from 60% – 90%. In order to detect frequencies this low with the branching process model, the model developed in Chapter 4 must be extended to specifically search for non-fixed loci. However, as previously mentioned this extended model is difficult to implement. Hence, if replicate data is available, these *t*-test methods should be used to analyse the marker frequencies rather than the branching process statistical model developed in Chapter 4.

References

Ayoub, M.Mather, D. E. (2002). Effectiveness of selective genotyping for detection of quantitative trait loci: an analysis of grain and malt quality traits in three barley populations. *Genome* **45,** 1116-24.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D.*, et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40,** 955-62.

Ben-Ari, G., Zenvirth, D., Sherman, A., David, L., Klutstein, M., Lavi, U.*, et al.* (2006). Four linked genes participate in controlling sporulation efficiency in budding yeast. *PLoS Genet* **2,** e195.

Brauer, M. J., Christianson, C. M., Pai, D. A. &Dunham, M. J. (2006). Mapping novel traits by array-assisted bulk segregant analysis in Saccharomyces cerevisiae. *Genetics* **173,** 1813-6.

Brockmann, G. A., Kratzsch, J., Haley, C. S., Renne, U., Schwerin, M. &Karle, S. (2000). Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F(2) variance of growth and obesity in DU6i x DBA/2 mice. *Genome Res* **10,** 1941-57.

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim (NY)* **30,** 44-52.

Carlborg, O.Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5,** 618-25.

Carlborg, O., Kerje, S., Schutz, K., Jacobsson, L., Jensen, P. &Andersson, L. (2003). A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res* **13,** 413-21.

Crow, J. F.Kimura, M. (1965). Evolution in sexual and asexual populations *Amer Natur* **99,** 439-450.

Culleton, R., Martinelli, A., Hunt, P. &Carter, R. (2005). Linkage group selection: rapid gene discovery in malaria parasites. *Genome Res* **15,** 92-7.

Darvasi, A. (1998). Experimental strategies for the genetic dissection of complex traits in animal models. *Nat Genet* **18,** 19-24.

Darvasi, A.Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* **85,** 353-359.

Darvasi, A.Soller, M. (1994). Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* **138,** 1365-73.

Deutschbauer, A. M.Davis, R. W. (2005). Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat Genet* **37,** 1333-40.

Dilda, C. L.Mackay, T. F. (2002). The genetic architecture of Drosophila sensory bristle number. *Genetics* **162,** 1655-74.

Ehrenreich, I. M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J. A., *et al.* (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464,** 1039-42.

Feller, W. (1951). Diffusion processes in genetics. *Proc. Second Berkeley Symp. Math. Statist. Prob***,** 227-246.

Flint, J.Mackay, T. F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res* **19,** 723-33.

Foolad, M. R., Zhang, L. P. &Lin, G. Y. (2001). Identification and validation of QTLs for salt tolerance during vegetative growth in tomato by selective genotyping. *Genome* **44,** 444-54.

Foolad, M. R., Zhang, L. P. &Subbiah, P. (2003). Genetics of drought tolerance during seed germination in tomato: inheritance and QTL mapping. *Genome* **46,** 536-45.

Graham, R. R., Kyogoku, C., Sigurdsson, S., Vlasova, I. A., Davies, L. R., Baechler, E. C., *et al.* (2007). Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci U S A* **104,** 6758-63.

Haley, C. S.Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69,** 315-24.

Jagers, P. (1975). *Branching processes with biological applications*. Wiley, London ; New York.

Keightley, P. D.Bulfield, G. (1993). Detection of quantitative trait loci from frequency changes of marker alleles under selection. *Genet Res* **62,** 195-203.

Kirkpatrick, B. W., Byla, B. M. &Gregory, K. E. (2000). Mapping quantitative trait loci for bovine ovulation rate. *Mamm Genome* **11,** 136-9.

Kroymann, J.Mitchell-Olds, T. (2005). Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435,** 95-8.

Lai, C. Q., Leips, J., Zou, W., Roberts, J. F., Wollenberg, K. R., Parnell, L. D., *et al.* (2007). Speed-mapping quantitative trait loci using microarrays. *Nat Methods* **4,** 839-41.

Lander, E. S.Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185-99.

Lebowitz, R. J., Soller, M. &Beckmann, J. S. (1987). Trait-Based Analyses for the Detection of Linkage between Marker Loci and Quantitative Trait Loci in Crosses between Inbred Lines. *Theoretical and Applied Genetics* **73,** 556-562.

Lynch, M.Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature* **461,** 747-53.

Martinelli, A., Cheesman, S., Hunt, P., Culleton, R., Raza, A., Mackinnon, M., *et al.* (2005). A genetic approach to the de novo identification of targets of strain-specific immunity in malaria parasites. *Proc Natl Acad Sci U S A* **102,** 814-9.

Michelmore, R. W., Paran, I. &Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect

markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci U S A* **88,** 9828-32.

Micic, Z., Hahn, V., Bauer, E., Melchinger, A. E., Knapp, S. J., Tang, S.*, et al.* (2005). Identification and validation of QTL for Sclerotinia midstalk rot resistance in sunflower by selective genotyping. *Theor Appl Genet* **111,** 233-42.

Nachman, M. W., Hoekstra, H. E. &D'Agostino, S. L. (2003). The genetic basis of adaptive melanism in pocket mice. *Proc Natl Acad Sci U S A* **100,** 5268-73.

Nones, K., Ledur, M. C., Ruy, D. C., Baron, E. E., Melo, C. M., Moura, A. S.*, et al.* (2006). Mapping QTLs on chicken chromosome 1 for performance and carcass traits in a broiler x layer cross. *Anim Genet* **37,** 95-100.

Nuzhdin, S. V., Harshman, L. G., Zhou, M. &Harmon, K. (2007). Genome-enabled hitchhiking mapping identifies QTLs for stress resistance in natural Drosophila. *Heredity* **99,** 313-21.

Nuzhdin, S. V., Keightley, P. D., Pasyukova, E. G. &Morozova, E. A. (1998). Mapping quantitative trait loci affecting sternopleural bristle number in Drosophila melanogaster using changes of marker allele frequencies in divergently selected lines. *Genet Res* **72,** 79-91.

Quarrie, S. A., Lazic-Jancic, V., Kovacevic, D., Steed, A. &Pekic, S. (1999). Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *Journal of Experimental Botany* **50,** 1299-1306.

Rice, W. R. (2002). Experimental tests of the adaptive significance of sexual recombination. *Nat Rev Genet* **3,** 241-51.

Ruyter-Spira, C. P., Gu, Z. L., Van der Poel, J. J. &Groenen, M. A. (1997). Bulked segregant analysis using microsatellites: mapping of the dominant white locus in the chicken. *Poult Sci* **76,** 386-91.

Sax, K. (1923). The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS. *Genetics* **8,** 552-60.

Segre, A. V., Murray, A. W. &Leu, J. Y. (2006). High-resolution mutation mapping reveals parallel experimental evolution in yeast. *PLoS Biol* **4,** e256.

Sinha, H., Nicholson, B. P., Steinmetz, L. M. &McCusker, J. H. (2006). Complex genetic interactions in a quantitative trait locus. *PLoS Genet* **2,** e13.

Steinmetz, L. M., Sinha, H., Richards, D. R., Spiegelman, J. I., Oefner, P. J., McCusker, J. H.*, et al.* (2002). Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416,** 326-30.

Thoday, J. M. (1961). Location of polygenes. *Nature* **191,** 368-370.

Vaughn, T. T., Pletscher, L. S., Peripato, A., King-Ellison, K., Adams, E., Erikson, C.*, et al.* (1999). Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. *Genet Res* **74,** 313-22.

Wenzl, P., Raman, H., Wang, J., Zhou, M., Huttner, E. &Kilian, A. (2007). A DArT platform for quantitative bulked segregant analysis. *BMC Genomics* **8,** 196.

Yamamoto, A., Anholt, R. R. &MacKay, T. F. (2009). Epistatic interactions attenuate mutations affecting startle behaviour in Drosophila melanogaster. *Genet Res* **91,** 373-82.

Zhang, L. P., Lin, G. Y., Nino-Liu, D. &Foolad, M. R. (2003). Mapping QTLs conferring early blight (Alternaria solani) resistance in a Lycopersicon

esculentumXL. hirsutum cross by selective genotyping. *Molecular Breeding* **12,** 3-19.